APPROXIMATION TECHNIQUES FOR INCOMPRESSIBLE FLOWS WITH HETEROGENEOUS PROPERTIES

A Dissertation

by

ABNER JONATAN SALGADO GONZALEZ

Submitted to the Office of Graduate Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

August 2010

Major Subject: Mathematics

APPROXIMATION TECHNIQUES FOR INCOMPRESSIBLE FLOWS WITH HETEROGENEOUS PROPERTIES

A Dissertation

by

ABNER JONATAN SALGADO GONZALEZ

Submitted to the Office of Graduate Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Approved by:

Chair of Committee,	Jean-Luc Guermond
Committee Members,	Raytcho D. Lazarov
	Bojan Popov
	Jim E. Morel
Head of Department,	Al Boggess

August 2010

Major Subject: Mathematics

ABSTRACT

Approximation Techniques for Incompressible Flows with Heterogeneous Properties.

(August 2010)

Abner Jonatan Salgado Gonzalez, B.S., Saint Petersburg State Polytechnic University;

M.S., Saint Petersburg State Polytechnic University

Chair of Advisory Committee: Dr. Jean-Luc Guermond

We study approximation techniques for incompressible flows with heterogeneous properties. Specifically, we study two types of phenomena. The first is the flow of a viscous incompressible fluid through a rigid porous medium, where the permeability of the medium depends on the pressure. The second is the flow of a viscous incompressible fluid with variable density. The heterogeneity is the permeability and the density, respectively.

For the first problem, we propose a finite element discretization and, in the case where the dependence on the pressure is bounded from above and below, we prove its convergence to the solution and propose an algorithm to solve the discrete system. In the case where the dependence is exponential, we propose a splitting scheme which involves solving only two linear systems.

For the second problem, we introduce a fractional time-stepping scheme which, as opposed to other existing techniques, requires only the solution of a Poisson equation for the determination of the pressure. This simplification greatly reduces the computational cost. We prove the stability of first and second order schemes, and provide error estimates for first order schemes.

For all the introduced discretization schemes we present numerical experiments, which illustrate their performance on model problems, as well as on realistic ones.

ACKNOWLEDGMENTS

I could have never been able to complete a work of this magnitude by myself. There is an enormous (but nevertheless finite) list of individuals, that in one way or another have contributed in getting me to this point. But, I am afraid that if I start writing names I might forget somebody. If that is the case I ask beforehand for their forgiveness.

First and foremost, I am very indebted to my parents. Their support has been invaluable through all these years.

I cannot omit mentioning my very first mentor, V.G. Korneev from St. Petersburg State Polytechnic University, who was the one who introduced me to the theory of finite elements and tried to teach me how to properly do numerical analysis. Without his guidance I would have never been able to come to Texas A&M University.

I want to thank my advisor, J.-L. Guermond, for all his patience, encouragement and interesting discussions that we have had through all these years. I could have never become the (hopefully successful) professional that I am today without his support and guidance. The interesting courses, conversations or comments that I have had from all the faculty members of the Numerical Analysis group have been crucial to my education, especially the ones from R. Lazarov, B. Popov, W. Bangerth and A. Bonito.

I am indebted to V. Girault for all her help, advice and support during the Summer of 2008. Without all her encouragement and guidance, practically half of this dissertation would have never been written. P. Minev has been very helpful in the late stages of my studies here at Texas A&M.

Last in this list of people, but by no means least, I want to thank all my good friends and fellow students (they know who they are). They all have encouraged me

to keep going when I needed it and they have made me stop when necessary as well. The sole knowledge that I can count on them on the good and bad times is more than sufficient for them to deserve a special mention in this list.

Finally, I would like to acknowledge the support from the Institute of Applied Mathematics and Computational Science through Award No. KUS-C1-016-04, made by King Abdullah University of Science and Technology (KAUST).

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

FIGURE Page

CHAPTER I

INTRODUCTION

The efficient and accurate numerical approximation of complicated fluid flow phenomena is of extreme importance for a wide range of applications. However, the complexity of the models that this requires poses serious challenges in various areas of mathematics. Just to mention a few, these areas might be the analysis of the mathematical models (equations) of these phenomena, trying to answer questions about well-posedness of these problems which, in some sense, is a minimal requirement for the consistency of a model. Another one is the development and analysis of efficient discretization schemes and solution techniques for these problems. Since as a rule these models are nonlinear, this always proves to be a highly nontrivial task.

The purpose of this dissertation is the study of effective discretization and solution techniques for problems that arise in the modeling of incompressible fluid flow that has heterogeneous properties. To be more precise, we will analyze two of these phenomena. The first one is related to the flow in porous media and a model that is used in the problem of enhanced oil recovery. It is a Darcy's model where the porosity of the medium depends on the pressure. The second problem is the flow of incompressible Newtonian fluids with variable density. This is a model that is frequently used in the study of multiphase flow, temperature dependent flow and others. In both cases, the flow has heterogeneous properties: the porosity and density, respectively. This heterogeneity highly complicates the model and the techniques that must be used to efficiently discretize and approximate the solution to them.

Let us briefly elaborate on each one of this models.

The journal model is SIAM Journal of Numerical Analysis.

A. Darcy's Equations with Pressure Dependent Porosity

The system of equations commonly referred to as Darcy's law was obtained, on the basis of experimental observations, by H. Darcy (cf. [23]) more than 150 years ago. This law approximates the balance of linear momentum of a fluid flow through a porous rigid body and is the simplest model of flow of a viscous incompressible fluid through a porous medium. Darcy's equations were obtained rigorously by Homogenization; without being exhaustive, we refer to the works of I.H. Ene and E. Sánchez-Palencia [26], G. Allaire [3], D. Cioranescu, P. Donato and I.H. Ene [22], S.E. Pastukhova [67], and E. Skjetne and J.-L. Auriault [77].

Recently, in [70], K.R. Rajagopal developed systematically a family of models within the framework of Mixture Theory, deriving first Darcy's system, and next relaxing one or more restrictions that were used in deriving this law. The steady nonlinear model studied in the present work is one of the numerous models obtained through this approach (cf. [70, Section 3.5]). It is a much simplified version of a model of enhanced oil recovery, where oil is forced to flow through rocks by injecting steam at high pressure. This model is simplified because only one fluid is considered and the viscous and inertial effects are neglected, thus resulting in a steady system. On the other hand, it is nonlinear because the porosity of the solid medium is allowed to depend exponentially on the pressure. Indeed, it has been observed experimentally that high variations on the pressure induce an exponential variation on the porosity of the medium.

Let Ω be a bounded domain in $\mathbb{R}^d$, with $d = 2, 3$. The boundary, $\partial\Omega$, of this domain is divided into two parts Γ_w and Γ. We are interested in the following model,

which as we have stated above was derived by K.R. Rajagopal [70],

$$\begin{cases} \alpha(p)\mathbf{u} + \nabla p = \mathbf{f}, & \text{in } \Omega, \\ \nabla\cdot\mathbf{u} = 0, & \text{in } \Omega, \\ p = 0, & \text{on } \Gamma_w, \\ \mathbf{u}\cdot\mathbf{n} = g & \text{on } \Gamma, \end{cases} \tag{1.1}$$

where the unknowns are the velocity $\mathbf{u}$ and the pressure p of the fluid. The function α is known as the *drag coefficient*, *permeability* or *inverse porosity*. It describes how easily the fluid can pass through the given medium, and for simplicity is assumed homogeneous.

In the case when α is constant or dependent only on the medium, these equations have been deeply studied, and the discretization techniques used in this case are well established. We refer, for instance, to [16], [1] or [27]. However, as it is noted in [70], experiments show that if the variations on the pressure are high, the material cracks and thus the porosity varies. For this reason, it is proposed to consider the case where the drag coefficient depends on the pressure. Moreover, the dependence that most accurately describes experimental phenomena near a well is an exponential one

$$\alpha(\xi) = \alpha_0 e^{\gamma\xi}, \tag{1.2}$$

for some positive parameters α_0, γ. The homogeneous boundary condition in the third row of (1.1) is just introduced to simplify the discussion. More generally, a non homogeneous boundary condition can be prescribed on the pressure: $p = p_w$ on Γ_w. Owing to the nature of $\alpha(p)$ the analysis we present readily carries over to this case for adequately smooth boundary data.

For the sake of brevity, in what follows we shall refer to equations (1.1) simply as

the nonlinear Darcy equations. Of course, there are other nonlinear Darcy's model, such as the well-known Forchheimer model introduced by Forchheimer in [30]. Concering its discretization, we refer to the study of a steady Forchheimer model studied by V. Girault and M.F. Wheeler in [35].

The analysis of the nonlinear Darcy equations is difficult because of the exponential nonlinearity. In this dissertation, following the work of M. Azaïez, F. Ben Belgacem, C. Bernardi, and N. Chorfi in [5], we propose first to discretize (1.1) when the function α is truncated above and below. We introduce a straightforward finite element scheme, such as $\mathbb{P}_{k-1}$ for each component of the velocity and $\mathbb{P}_k$ for the pressure, similar to the scheme studied by J.E. Roberts and J.-M. Thomas in [72] and by D. Kim and E.J. Park in [59]. When the exact solution is sufficiently small so that it satisfies a sufficient condition for uniqueness, we establish optimal a priori error estimates, and geometric convergence of a successive approximation algorithm for computing the discrete solution. We also study the case when the exact solution is nonsingular in the sense of F. Brezzi, J. Rappaz and P.-A. Raviart [17], but is not necessarily unique. We give sufficient conditions for the finite element scheme to have a nonsingular solution, establish convergence and a priori error estimates, and study the convergence of Newton's algorithm for computing this solution. In particular, we prove that Newton's method converges quadratically, but not uniformly. This confirms the convergence analysis for nonlinear second order elliptic problems studied by J. Douglas and T. Dupont in [24] and by E.J. Park in [66].

Next, we study the problem with fully exponential porosity. To begin with, the velocity is eliminated by:

1. dividing the equation by the exponential,

2. taking the divergence of the equation,

3. and making a change in variable.

This splits the problem into exactly two consecutive *linear* equations: first a diffusion–convection–reaction equation and next a linear Darcy system. These are discretized by an easy variant of the finite element scheme used in the first approach. The analysis of each discrete linear system is straightforward, but the global analysis of the complete algorithm is still an open problem.

B. The Variable Density Navier-Stokes Equations

The flow of incompressible viscous fluids with variable density, under certain assumptions, is governed by the time-dependent Navier-Stokes equations:

$$\begin{cases} \rho_t + \nabla\cdot(\rho \mathsf{u}) = 0, \\ \rho(\mathsf{u}_t + \mathsf{u}\cdot\nabla\mathsf{u}) + \nabla \mathsf{p} - \mu\Delta\mathsf{u} = \mathbf{f}, \\ \nabla\cdot\mathsf{u} = 0, \end{cases} \tag{1.3}$$

where the unknowns are the density $\rho > 0$, the velocity field u, and the pressure p. The constant μ is the dynamic viscosity coefficient and $\mathbf{f}$ is a driving external force. In stratified flows we typically have $\mathbf{f} = \rho\mathbf{g}$, where $\mathbf{g}$ is the gravity field. The fluid occupies a bounded domain Ω in $\mathbb{R}^d$ (with $d = 2$ or 3) and a solution to the above problem is sought over a time interval $[0, T]$. The Navier-Stokes system is supplemented by the following initial and boundary conditions for u and ρ:

$$\begin{cases} \rho(x,0) = \rho_0(x), \quad \rho(x,t)|_{\Gamma^-} = a(\mathbf{x},t), \\ \mathsf{u}(\mathbf{x},0) = \mathbf{u}_0(\mathbf{x}), \quad \mathsf{u}(\mathbf{x},t)|_{\partial\Omega} = \mathbf{b}(\mathbf{x},t), \end{cases} \tag{1.4}$$

Γ^- is the inflow boundary, which is defined by

$$\Gamma^- = \{\mathbf{x} \in \Gamma : \ \mathsf{u}(\mathbf{x})\cdot\mathbf{n} < 0\},$$

with $\mathbf{n}$ being the outward unit normal vector. Throughout this dissertation we assume that the boundary Γ is impermeable, i.e., $\mathsf{u}\cdot\mathbf{n} = 0$ everywhere on Γ, and $\Gamma^- = \emptyset$.

The mathematical theory of existence and uniqueness for (1.3)–(1.4) is quite involved and far from complete. We refer the reader to the works of P.L. Lions [61], E. Fernández-Cara and F. Guillén [28] for further details. The difficulty comes from the fact that these equations entangle hyperbolic, parabolic, and elliptic features. Approximating (1.3)–(1.4) efficiently is a challenging task. A testimony of the difficulty is that, so far, very few papers have been dedicated to the mathematical analysis of the approximation of (1.3)–(1.4). We refer to C.L. Liu and N.J. Walkington [63] for one of the few attempts in this direction.

Approximating (1.3)–(1.4) can be done by solving the coupled system (1.3), but this approach may sometimes be computer intensive due to saddle point structure that the incompressibility induces in the problem. Alternative, more efficient, approaches advocated in the literature consist of using fractional time-stepping and exploiting, as far as possible, techniques already established for the solution of constant density incompressible fluid flows. The starting point of most fractional time-stepping algorithms consists of decoupling the incompressibility constraint and diffusion in the spirit of A.J. Chorin's [20] and R. Temam's [79] projection method. Several algorithms have been developed which extend this idea to the case of variable density flows, see for example J.B. Bell and D.L. Marcus [11], A. Almgren et al. [4], J.-L. Guermond and L. Quartapelle [48], and J.-H. Pyo and J. Shen [69]. To the best of our knowledge, [48] gave the first stability proof of a projection method for variable density flows. The algorithm proposed in [48] is somewhat expensive since it is composed of two time-consuming projections. An alternative algorithm composed of only one projection per time step was proposed in [69] and proved to be stable. It seems that so far [48] and [69] are the only papers where projection methods for variable density flows

have been proved to be stable, the best available results being that of [69]. However, no rigorous error analysis of these methods is available in the literature.

The common feature of all the projection-like methods referred to above is that at each time step, say t^{n+1}, the pressure or some related scalar quantity, say Φ, is determined by solving an equation of the following form:

$$-\nabla\cdot\left(\frac{1}{\rho^{k+1}}\nabla\Phi\right)=\Psi,\qquad \partial_n\Phi|_\Gamma=0, \tag{1.5}$$

where ρ^{k+1} is an approximation of the density at time t_{k+1} and Ψ is some right-hand side that varies at each time step. The problem (1.5) is far more complicated to solve than just a Poisson equation. It is time consuming since it requires assembling and pre-conditioning a variable-coefficient stiffness matrix at each time step. Note also in passing that it is necessary to have a uniform lower bound on the value of the density for (1.5) to be solvable. This condition is often overlooked in the literature.

On the basis of the observations above, in this dissertation we introduce a family of fractional time-stepping methods for solving variable density flows that involve solving only one Poisson problem per time step instead of problems like (1.5). We will show the stability and convergence properties of the first order schemes and the stability of a formally second order variant.

CHAPTER II

PRELIMINARIES

The purpose of this chapter is to establish the notation that shall be used in the subsequent chapters. In the following, we denote by c a generic constant, the value of which may vary at each occurrence. When studying continuous problems, the value of this constant may depend on the data of the problem, but not on the solution. On the other hand, when studying the discretization of a problem, the value of this constant may depend on the data of a problem and its exact solution, but it does not depend on the discretization parameters or the solution of the numerical scheme.

A. Function Spaces

Henceforth, we denote by Ω a bounded connected domain in $\mathbb{R}^d$, with $d = 2$ or 3. The boundary of this domain is denoted by $\partial\Omega$. As usual, we denote by $L^q(\Omega)$ the space of Lebesgue integrable functions with exponent $q \in [1, \infty]$ defined on Ω and normed, for $1 \leq q < \infty$, by

$$\|v\|_{L^q} := \left(\int_\Omega |v|^q \right)^{1/q},$$

and, for $q = \infty$

$$\|v\|_{L^\infty} := \operatorname*{esssup}_{\mathbf{x} \in \Omega} |v|.$$

For which these spaces are Banach spaces. In the case $q = 2$ we denote by $\langle \cdot, \cdot \rangle$ the L^2-scalar product.

By $W_q^s(\Omega)$, for an integer s, we denote the Sobolev space of functions in $L^q(\Omega)$ with partial derivatives of order up to s in $L^q(\Omega)$, namely

$$W_q^s(\Omega) := \left\{ v \in L^q(\Omega) : \ \partial^m v \in L^q(\Omega), \forall |m| \leq s \right\},$$

equipped with the seminorm

$$|v|_{W_q^s} := \left(\sum_{|m|=s} \int_\Omega |\partial^m v|^q \right)^{1/q},$$

and norm (for which it is a Banach space)

$$\|v\|_{W_q^l} := \left(\sum_{0 \le |m| \le s} |w|^q_{W_q^m} \right)^{1/q}.$$

When s is not an integer, $W_p^s(\Omega)$ is defined using the real method of interpolation (cf. J.L Lions and E. Magenes [60] or J. Berg and J. Löfstrom [12]). In this case, there are several equivalent norms. Here, we choose the following seminorm and norm: let $s = m + s'$ for an integer $m \ge 0$ and $0 < s' < 1$, then we set

$$|v|_{W_q^s} := \left(\sum_{|l|=m} \int_\Omega \int_\Omega \frac{|\partial^l v(\mathbf{x}) - \partial^l v(\mathbf{y})|^q}{|\mathbf{x}-\mathbf{y}|^{d+qs'}} \right)^{1/q},$$

$$\|v\|_{W_q^s} := \left(\|v\|^q_{W_q^m} + |v|_{W_q^s} \right)^{1/q}.$$

When $q = 2$ we set $H^s(\Omega) := W_2^s(\Omega)$ for any s. By $H_0^1(\Omega)$ we denote the closure of $\mathcal{C}_0^\infty(\Omega)$ in the H^1-norm.

In Chapter III the following trace property will be needed. If the domain Ω has a Lipschitz-continuous boundary and v belongs to $H^s(\Omega)$ for $s \in (1/2, 1]$ then it has a well defined trace on the boundary, this trace belongs to $H^{s-1/2}(\partial\Omega)$ (cf. P. Grisvard [36, Theorem 1.5.1.2]) and

$$\|v\|_{H^{s-1/2}} \le c \|v\|_{H^s}.$$

In this chapter, the space $H_{00}^{1/2}(\Gamma)$ will also be needed, this space is defined as follows. Let Γ be a subset of $\partial\Omega$ that has positive measure, we say that a function $g \in H^{1/2}(\Gamma)$ belongs to $H_{00}^{1/2}(\Gamma)$ if its extension by zero to $\partial\Omega$ belongs to $H^{1/2}(\partial\Omega)$. For a discussion

on this space see L. Tartar [78], for instance.

There are several well known embedding theorems for Sobolev spaces. We shall use repeatedly the embedding $H^1(\Omega) \hookrightarrow L^6(\Omega)$ which, given enough smoothness of the domain, is valid for $d \leq 3$ (cf. R.A. Adams [2] or [78]). When we wish to indicate explicitly that we are using the constant of this embedding, we denote it by $c(\Omega)$. That is, by $c(\Omega)$ we denote the smallest constant such that

$$\|q\|_{L^6} \leq c(\Omega)|q|_{H^1}, \quad \forall q \in H^1(\Omega).$$

Finally, we must state that we use bold-face characters to denote vector valued functions and their spaces.

B. Time Dependent Problems

Chapter IV is dedicated to the study of a time dependent problem. Here we introduce some notation that shall be used in this chapter.

Whenever E is a normed space with norm $\|\cdot\|_E$, we say that a function $\phi : [0,T] \to E$ belongs to $L^q(0,T;E)$ (which will also be denoted by $L^q(E)$) if the map $(0,T) \ni t \mapsto \|\phi(t)\|_E$ is L^q integrable. A similar definition allows us to define the spaces $W^s_q(E)$.

When introducing a time discretization, we denote by $\tau > 0$ a time step and we set $t_k = k\tau$ for $0 \leq k \leq K := [T/\tau]$. For any time-dependent function $\phi : [0,T] \to E$, we denote by $\phi^k := \phi(t_k)$. The sequence $\phi^0, \phi^1, \ldots, \phi^K$ is denoted ϕ_τ. To shorten the notation, we introduce the time-increment operator δ by setting

$$\delta\phi^k = \phi^k - \phi^{k-1}.$$

Finally, the errors of our discretization schemes will be measured in the following

discrete norms:

$$\|\phi_\tau\|_{\ell^2(E)} := \left(\tau \sum_{k=0}^{K} \|\phi^k\|_E^2 \right)^{1/2}, \qquad \|\phi_\tau\|_{\ell^\infty(E)} := \max_{0 \le k \le K} \left(\|\phi^k\|_E \right).$$

Which, clearly, are consistent with the $L^2(E)$ and $L^\infty(E)$, respectively, as $\tau \to 0$.

CHAPTER III

NONLINEAR DARCY EQUATIONS *

In this chapter we study problem (1.1). The results of this chapter were originally presented in [32], and the organization is as follows. In Section A we study the mathematical analysis of the problem, i.e., questions regarding the existence and uniqueness (both global and local) of a solution to this problem. In Section B we analyze the discretization of this problem in the case when the porosity is uniformly bounded from above and below. We present discretization schemes for the case when the solution is unique and non-singular. Section C is dedicated to the case of an exponential porosity and proposes a solution scheme for this case. Finally, Section D presents some numerical experiments that illustrate the algorithms introduced in the previous sections.

A. Analysis of the Problem

Before considering the discretization of problem (1.1) we will discuss some properties of its exact solution, namely its existence and sufficient conditions for this solution to be globally unique and possess certain smoothness properties. When the nonlinear Darcy equations have more than one solution we shall discuss the so-called *nonsingular solutions*, in the sense of [17]. This shall prove useful for the development and analysis of the discretization.

We intend to study problem (1.1) under the following assumptions:

- The domain Ω has a Lipschitz-continuous boundary $\partial\Omega$ divided into two parts Γ_w and Γ, also with Lipschitz continuous boundaries.
- The part of the boundary Γ_w has positive surface measure.
- The function $\alpha : \mathbb{R} \to \mathbb{R}$ is continuous and there are two positive constants $\alpha_{\min}$ and $\alpha_{\max}$ such that

$$\alpha_{\min} \leq \alpha(\xi) \leq \alpha_{\max}, \quad \forall \xi \in \mathbb{R}. \tag{3.1}$$

- The function α is uniformly Lipschitz-continuous on $\mathbb{R}$. That is, there is a constant $L_\alpha > 0$ such that for all $\xi_1, \xi_2 \in \mathbb{R}$

$$|\alpha(\xi_1) - \alpha(\xi_2)| \leq L_\alpha |\xi_1 - \xi_2| . \tag{3.2}$$

Remark 1. Assumptions (3.1) and (3.2) are not true when the function α is unbounded, as it is the case when it is exponential. However, these assumptions can be easily recovered by truncating the original function α. Obviously, the solution of the truncated problem will not in general solve the original one. The analysis of how these two problems are related is beyond the scope of this work.

It is well known that Darcy's equations have several variational formulations. We have chosen here the formulation that treats the boundary condition on p as an essential one and leads, roughly speaking, to taking $\mathbf{u}$ in $\mathbf{L}^2(\Omega)$ and p in $H^1(\Omega)$. This choice is motivated by the fact that the forthcoming analysis of the nonlinear term $\alpha(p)\mathbf{u}$ uses intensively the fact that p is in $H^1(\Omega)$. Moreover, a velocity $\mathbf{u}$ in $\mathbf{L}^2(\Omega)$ is easily discretized. Another option consists in taking $\mathbf{u}$ in $\mathbf{H}(\mathrm{div};\Omega)$ and p in $L^2(\Omega)$. Then $\mathbf{u}$ must be discretized with mixed finite elements, with the advantage that this leads to a locally conservative scheme. But the drawback is that the analysis of the nonlinear term is not so clear.

Let us define the space

$$H^1_w(\Omega) := \left\{ q \in H^1(\Omega) : \, q|_{\Gamma_w} = 0 \right\},$$

and assume, for the sake of simplicity, that $p_w = 0$. Then the variational formulation is the following:

Given $\mathbf{f} \in \mathbf{L}^2(\Omega)$ *and* $g \in H^{1/2}_{00}(\Gamma)'$, *find a pair* $(\mathbf{u}, p) \in \mathbf{L}^2(\Omega) \times H^1_w(\Omega)$ *such that*

$$\begin{cases} a(p; \mathbf{u}, \mathbf{v}) + b(\mathbf{v}, q) = \int_\Omega \mathbf{f} \cdot \mathbf{v}, & \forall \mathbf{v} \in \mathbf{L}^2(\Omega), \\ b(\mathbf{u}, q) = \langle g, q \rangle_\Gamma, & \forall q \in H^1_w(\Omega). \end{cases} \tag{3.3}$$

The bilinear forms $a(\xi; \cdot, \cdot)$ for any measurable function ξ on Ω and $b(\cdot, \cdot)$ are defined by

$$a(\xi; \mathbf{v}, \mathbf{w}) := \int_\Omega \alpha(\xi) \mathbf{v} \cdot \mathbf{w}, \tag{3.4}$$

$$b(\mathbf{v}, q) := \int_\Omega \mathbf{v} \cdot \nabla q, \tag{3.5}$$

and $\langle \cdot, \cdot \rangle_\Gamma$ denotes the duality pairing between $H^{1/2}_{00}(\Gamma)$ and its dual space $H^{1/2}_{00}(\Gamma)'$.

It is readily checked that under assumption (3.1) the forms $a(\xi; \cdot, \cdot)$ and $b(\cdot, \cdot)$ are continuous on $\mathbf{L}^2(\Omega) \times \mathbf{L}^2(\Omega)$ and $\mathbf{L}^2(\Omega) \times H^1(\Omega)$ respectively. Thus, standard arguments yield the equivalence of problem (3.3) with the system (1.1) in the distribution sense.

Remark 2. The above variational formulation is defined for homogeneous boundary conditions: $p_w = 0$. Standard techniques (i.e., lifting arguments) allow us to reduce the case of nonhomogeneous Dirichlet boundary conditions on the pressure p to the present one. For this, it is sufficient to assume that $p_w \in H^{1/2}(\Gamma_w)$ and notice that the function $\xi \mapsto \alpha(\xi - \bar{p}_w)$, where $\bar{p}_w$ is a proper lifting of p_w, has the same properties as $\xi \mapsto \alpha(\xi)$. Hence, there is no loss of generality in considering only homogeneous

Dirichlet boundary conditions.

The existence of a solution to problem (3.3) is studied in [5]. For the sake of completeness we list here the results that later prove useful for our purposes. Regarding existence we have the following Theorem.

Theorem 1. *Assume that the function α satisfies assumption* (3.1). *Then, for any data $(\mathbf{f}, g) \in \mathbf{L}^2(\Omega) \times H_{00}^{1/2}(\Gamma)'$ problem* (3.3) *has a solution $(\mathbf{u}, p) \in \mathbf{L}^2(\Omega) \times H_w^1(\Omega)$. Moreover, this solution satisfies*

$$\|\mathbf{u}\|_{\mathbf{L}^2} + \|p\|_{H^1} \leq c \left(\|\mathbf{f}\|_{\mathbf{L}^2} + \|g\|_{(H_{00}^{1/2})'} \right). \tag{3.6}$$

A sufficient condition for the global uniqueness of the solution is given by the following Proposition.

Proposition 1. *Assume that the function α satisfies assumptions* (3.1) *and* (3.2). *If problem* (3.3) *has a solution $(\mathbf{u}, p)$ such that $\mathbf{u} \in \mathbf{L}^r(\Omega)$ with $r > d$, where d is the space dimension, and satisfies*

$$\frac{\alpha_{\max} + \alpha_{\min}}{\alpha_{\min}} c(r, \Omega) L_\alpha \|\mathbf{u}\|_{\mathbf{L}^3} < 1. \tag{3.7}$$

for an appropriate constant $c(r, \Omega)$ that depends only on r and Ω. Then, there is no other solution to problem (3.3).

Remark 3. Examining the proof given in [5] we see that the constant $c(r, \Omega)$ in the smallness condition (3.7) is the norm of the Sobolev embedding $H_w^1(\Omega) \hookrightarrow L^{r'}(\Omega)$ with $\frac{1}{r} + \frac{1}{r'} = \frac{1}{2}$. Moreover, the condition $r > d$ is due to the Sobolev embedding when $d = 2$. However, when $d = 3$, this proof is also valid with $r = 3$. For the sake of definiteness, in the sequel, we shall assume that $d = 3$. The reader can verify that similar arguments, and less restrictive assumptions, yield the results for $d = 2$.

Finally, concerning the regularity of the solution the following result holds.

Proposition 2. *There exists a real number $\rho_0 > 2$ only depending on the geometry of Ω such that, for all ρ such that $2 < \rho \leq \rho_0$, and for all data $(\mathbf{f}, g) \in \mathbf{L}^\rho(\Omega) \times W_\rho^{-1/\rho}(\Gamma)$, any solution $(\mathbf{u}, p)$ to problem (3.3) belongs to $\mathbf{L}^\rho(\Omega) \times W_\rho^1(\Omega)$.*

Remark 4. The existence of ρ_0 is obtained in [5] by a perturbation argument, but in dimension $d = 3$, there is no guarantee that $\rho_0 \geq 3$. Therefore, in general, condition (3.7) for global uniqueness cannot be checked from the data.

Let us now consider the case when the solution is only locally unique. In this case, although problem (3.3) may have more than one solution, we assume that there exists an *isolated* solution. That is, there exists a neighborhood of this solution where no other solution exists. A sufficient condition for this to hold is that the solution is *nonsingular* (cf. [17] or V. Girault and P.-A. Raviart [34]). We shall analyze the properties of nonsingular solutions, and give sufficient conditions for such a solution to exist.

First we cast problem (3.3) in a more convenient, but nevertheless equivalent, functional setting. With this purpose let us define the data space

$$\mathfrak{Y} := \mathbf{L}^2(\Omega) \times H_{00}^{1/2}(\Gamma)',$$

with norm

$$\|(\mathbf{f}, g)\|_{\mathfrak{Y}} := \|\mathbf{f}\|_{\mathbf{L}^2} + \|g\|_{(H_{00}^{1/2})'},$$

and the solution space

$$\mathfrak{X} := \mathbf{L}^2(\Omega) \times H_w^1(\Omega),$$

with norm

$$\|(\mathbf{u}, p)\|_{\mathfrak{X}} := \|\mathbf{u}\|_{\mathbf{L}^2} + \|p\|_{H^1}.$$

We also define T as the solution operator to the linear Darcy problem, i.e., $T : \mathfrak{Y} \to \mathfrak{X}$

is such that, for every $\eta = (\mathbf{f}, g) \in \mathfrak{Y}$, $\mathfrak{X} \ni x = (\mathbf{u}, p) = T\eta = T(\mathbf{f}, g)$ solves

$$\begin{cases} \bar{\alpha}\mathbf{u} + \nabla p = \mathbf{f}, & \text{in } \Omega, \\ \nabla\cdot\mathbf{u} = 0, & \text{in } \Omega, \\ p = 0, & \text{on } \Gamma_w, \\ \mathbf{u}\cdot\mathbf{n} = g, & \text{on } \Gamma, \end{cases} \tag{3.8}$$

for a fixed $\bar{\alpha} > 0$.

It is classical that problem (3.8) is well-posed. This implies that $T \in \mathcal{L}(\mathfrak{Y}, \mathfrak{X})$. In other words, there is a constant $c > 0$ such that for every $(\mathbf{f}, g) \in \mathfrak{Y}$

$$\|T(\mathbf{f}, g)\|_{\mathfrak{X}} \leq c \, \|(\mathbf{f}, g)\|_{\mathfrak{Y}} . \tag{3.9}$$

By assumption (3.1) we get that $\alpha \in L^\infty(\mathbb{R})$. Then, for any $(\mathbf{u}, p) \in \mathfrak{X}$ we can conclude that $\alpha(p)\mathbf{u}$ is in $\mathbf{L}^2(\Omega)$ and we can define the map $G : \mathfrak{X} \to \mathfrak{Y}$ as follows. If $x = (\mathbf{u}, p)$ is an element of $\mathfrak{X}$, then

$$G(x) := \begin{pmatrix} (\alpha(p) - \bar{\alpha})\,\mathbf{u} - \mathbf{f} \\ -g \end{pmatrix} \in \mathfrak{Y}.$$

Finally, let us define $F : \mathfrak{X} \to \mathfrak{X}$ as

$$F(x) := x + TG(x).$$

With this notation, problem (3.3) can be equivalently restated as:

Find $x = (\mathbf{u}, p) \in \mathfrak{X}$ *such that*

$$F(x) = 0. \tag{3.10}$$

We are now in a position to define the notion of nonsingular solutions

Definition 1 ([17])**.** Let $x \in \mathfrak{X}$ solve problem (3.10). This solution is called *nonsingular* if the linear operator

$$F'(x) = I + TG'(x),$$

is an isomorphism of $\mathfrak{X}$. Here $F'(x)$ and $G'(x)$ denote the Fréchet derivative of the maps F and G at the point x, respectively.

Let us now provide sufficient conditions for a solution to be nonsingular in this sense. With this in mind, first of all, by assumption (3.2) we know that the derivative of α exists a.e. on $\mathbb{R}$ (cf. G.B. Folland [29]). Denoting this derivative by $\dot{\alpha}$ we can, formally, obtain the derivative of the map G. Let $x = (\mathbf{u}, p),\ y = (\mathbf{v}, q) \in \mathfrak{X}$, then

$$G'(x)y = \begin{pmatrix} (\alpha(p) - \bar{\alpha})\,\mathbf{v} + \dot{\alpha}(p)q\mathbf{u} \\ 0 \end{pmatrix}. \tag{3.11}$$

From this we can conclude that if $x = (\mathbf{u}, p) \in \mathbf{L}^3(\Omega) \times H^1_w(\Omega) \subset \mathfrak{X}$, the Fréchet derivative of the map G is well-defined, given by equation (3.11), and $G'(x) \in \mathcal{L}(\mathfrak{X}, \mathfrak{Y})$.

Remark 5. In the case $d = 3$, we need $\mathbf{u} \in \mathbf{L}^3(\Omega)$ because of the term $\dot{\alpha}(p)q\mathbf{u}$. Indeed, by assumption (3.2), Hölder's inequality and the Sobolev embedding $H^1 \hookrightarrow L^6$, we have

$$\int_\Omega |\dot{\alpha}(p)q\mathbf{u}|^2 \le L_\alpha^2 \left(\int_\Omega q^6\right)^{1/3} \left(\int_\Omega |\mathbf{u}|^3\right)^{2/3} \le c(\Omega)^2 L_\alpha^2 \|q\|^2_{H^1} \|\mathbf{u}\|^2_{\mathbf{L}^3},$$

where all inequalities are sharp. Clearly, if $d = 2$ we should require $\mathbf{u} \in \mathbf{L}^{2+\epsilon}(\Omega)$ for some $\epsilon > 0$. In both cases, we must assume that the velocity $\mathbf{u}$ lies in a smaller space than $\mathbf{L}^2(\Omega)$ for the derivative to make sense. This is in contrast to the common feature of many nonlinear operators arising in the analysis of partial differential equations that describe physical phenomena. For such an operator, its derivative is everywhere defined and the range of the derivative is a smaller space (i.e., smoother or more regular) than the data space. For this reason, we say that the operator G does

not have *regularizing properties*. The fact that for problem (1.1) the nonlinearity G does not have regularizing properties lies at the heart of all the difficulties that its theoretical and numerical analysis present.

We now give sufficient conditions for a solution of problem (3.10) to be nonsingular in the sense of Definition 1.

Proposition 3. *Assume that for problem* (3.10) *the function* α *is such that conditions* (3.1) *and* (3.2) *hold. Let* $x = (\mathbf{u}, p) \in \mathfrak{X}$ *be a solution to problem* (3.10). *If* $\mathbf{u} \in \mathbf{L}^3(\Omega)$ *and*

$$\frac{\alpha_{\max} + \alpha_{\min}}{\alpha_{\min}} c(\Omega) L_\alpha \|\mathbf{u}\|_{\mathbf{L}^3} < 1, \tag{3.12}$$

then this solution is nonsingular.

Proof. We need to show that the map $I + TG'(x)$ is an isomorphism of $\mathfrak{X}$. Since the operator is continuous, by the Open Mapping Theorem (cf. A.Ya. Helemskii [54]) it is sufficient to show that the operator is bijective. That is, given any $z = (\mathbf{w}, r) \in \mathfrak{X}$ there exists a unique $y = (\mathbf{v}, q) \in \mathfrak{X}$ such that

$$y + TG'(x)y = z,$$

or

$$(y - z) = T(-G'(x))y.$$

In other words, we must prove that the problem: Find $(\mathbf{v}, q) \in \mathfrak{X}$ such that

$$\begin{cases} \bar{\alpha}(\mathbf{v} - \mathbf{w}) + \nabla(q - r) = (\bar{\alpha} - \alpha(p))\,\mathbf{v} - \dot{\alpha}(p) q \mathbf{u}, & \text{in } \Omega, \\ \nabla\cdot(\mathbf{v} - \mathbf{w}) = 0, & \text{in } \Omega, \\ (\mathbf{v} - \mathbf{w})\cdot\mathbf{n} = 0, & \text{on } \Gamma, \\ q - r = 0, & \text{on } \Gamma_w, \end{cases}$$

always has a unique solution. Doing the elementary change of variables $(\mathbf{V}, Q) = (\mathbf{v} - \mathbf{w}, q - r) \in \mathfrak{X}$ this problem can be equivalently restated as: Find $(\mathbf{V}, Q) \in \mathfrak{X}$ such that

$$\begin{cases} \alpha(p)\mathbf{V} + \nabla Q = \mathbf{F}(Q), & \text{in } \Omega, \\ \nabla\cdot\mathbf{V} = 0, & \text{in } \Omega, \\ \mathbf{V}\cdot\mathbf{n} = 0, & \text{on } \Gamma, \\ Q = 0, & \text{on } \Gamma_w, \end{cases}$$

where

$$\mathbf{F}(Q) := (\bar{\alpha} - \alpha(p))\mathbf{w} - \dot{\alpha}(p)r\mathbf{u} - \dot{\alpha}(p)Q\mathbf{u} = \mathbf{F} + \bar{\mathbf{F}}(Q),$$

with

$$\mathbf{F} = (\bar{\alpha} - \alpha(p))\mathbf{w} - \dot{\alpha}(p)r\mathbf{u}, \quad \bar{\mathbf{F}}(Q) = \dot{\alpha}(p)Q\mathbf{u}.$$

Notice that, since $\mathbf{u} \in \mathbf{L}^3(\Omega)$ then $\mathbf{F}(Q) \in \mathbf{L}^2(\Omega)$. This problem can be written in variational form as: Find $(\mathbf{V}, Q) \in \mathfrak{X}$ such that

$$\begin{cases} \int_\Omega \alpha(p)\mathbf{V}\cdot\mathbf{W} + \int_\Omega \mathbf{W}\cdot\nabla Q = \int_\Omega \mathbf{F}(Q)\cdot\mathbf{W}, & \forall \mathbf{W} \in \mathbf{L}^2(\Omega), \\ \int_\Omega \mathbf{V}\cdot\nabla R = 0, & \forall R \in H^1_w(\Omega). \end{cases} \tag{3.13}$$

We observe that (3.13) is a linear Darcy's system with an affine perturbation $\mathbf{F}(Q)$. If we define the bilinear form $\mathcal{A} : \mathfrak{X} \times \mathfrak{X} \to \mathbb{R}$ by

$$\mathcal{A}\left[(\mathbf{V}, Q), (\mathbf{W}, R)\right] := \int_\Omega \alpha(p)\mathbf{V}\cdot\mathbf{W} + \int_\Omega \mathbf{W}\cdot\nabla Q + \int_\Omega \mathbf{V}\cdot\nabla R,$$

and assume for the moment that $\bar{\mathbf{F}}(Q) = 0$, i.e., $\mathbf{F}(Q)$ does not depend on Q, then, problem (3.13) has a unique solution if and only if:

1. There exists a constant $\beta_{\mathcal{A}} > 0$ such that

$$\inf_{0\neq(\mathbf{V},Q)\in\mathfrak{X}} \sup_{0\neq(\mathbf{W},R)\in\mathfrak{X}} \frac{\mathcal{A}\left[(\mathbf{V}, Q), (\mathbf{W}, R)\right]}{\|(\mathbf{V}, Q)\|_{\mathfrak{X}}\|(\mathbf{W}, R)\|_{\mathfrak{X}}} \geq \beta_{\mathcal{A}}. \tag{3.14}$$

2. The form $\mathcal{A}$ has the following property:

$$(\mathcal{A}\left[(\mathbf{V},Q),(\mathbf{W},R)\right]=0\ \forall(\mathbf{V},Q)\in\mathfrak{X})\Rightarrow(\mathbf{W},R)=0. \tag{3.15}$$

These two properties are equivalent to the fact that the linear Darcy problem defined by the form $\mathcal{A}$ is well-posed, which is a classical result. This also implies the *a priori* estimate

$$\|\mathbf{V}\|_{\mathbf{L}^2}+|Q|_{H^1}\leq c\,\|\mathbf{F}\|_{\mathbf{L}^2}\,, \tag{3.16}$$

for some $c>0$ that does not depend on $\mathbf{F}$, $\mathbf{V}$ or Q. Now, the well-posedness of (3.13) follows immediately by proving that the affine mapping $S\mapsto Q$, where Q is the second component of the solution pair $(\mathbf{V},Q)$ of (3.14) with data $\mathbf{F}(S)$ is a contraction, i.e., there exists $\mathcal{K}\in(0,1)$ such that

$$|Q|_{H^1}\leq\mathcal{K}|S|_{H^1},\quad\forall S\in H^1(\Omega).$$

To do this, let S be given in $H^1(\Omega)$, set $\mathbf{F}=0$, and take $\mathbf{W}=\mathbf{V}$ in the first equation of problem (3.13). The second equation, together with condition (3.1) imply

$$\alpha_{\min}\|\mathbf{V}\|^2_{\mathbf{L}^2}\leq\int_\Omega\alpha(p)\mathbf{V}\cdot\mathbf{V}=\int_\Omega\bar{\mathbf{F}}(S)\cdot\mathbf{V}\leq\|\bar{\mathbf{F}}(S)\|_{\mathbf{L}^2}\|\mathbf{V}\|_{\mathbf{L}^2},$$

or

$$\|\mathbf{V}\|_{\mathbf{L}^2}\leq\frac{1}{\alpha_{\min}}\|\bar{\mathbf{F}}(S)\|_{\mathbf{L}^2}.$$

By taking $\mathbf{W}=\nabla Q$ we obtain

$$\begin{aligned}|Q|^2_{H^1}&=\int_\Omega\nabla Q\cdot\nabla Q=\int_\Omega\bar{\mathbf{F}}(S)\cdot\nabla Q-\int_\Omega\alpha(p)\mathbf{V}\cdot\nabla Q\\&\leq\|\bar{\mathbf{F}}(S)\|_{\mathbf{L}^2}|Q|_{H^1}+\alpha_{\max}\|\mathbf{V}\|_{\mathbf{L}^2}|Q|_{H^1}\\&\leq\left(1+\frac{\alpha_{\max}}{\alpha_{\min}}\right)\|\bar{\mathbf{F}}(S)\|_{\mathbf{L}^2}|Q|_{H^1}\leq\left(1+\frac{\alpha_{\max}}{\alpha_{\min}}\right)\|\dot{\alpha}(p)S\mathbf{u}\|_{\mathbf{L}^2}|Q|_{H^1}.\end{aligned}$$

Since

$$\|\dot{\alpha}(p)S\mathbf{u}\|_{\mathbf{L}^2} \leq c(\Omega)L_\alpha\|\mathbf{u}\|_{\mathbf{L}^3}|S|_{H^1},$$

we derive

$$|Q|_{H^1} \leq \left(1+\frac{\alpha_{\max}}{\alpha_{\min}}\right)c(\Omega)L_\alpha\|\mathbf{u}\|_{\mathbf{L}^3}|S|_{H^1}.$$

Therefore the mapping $S \mapsto Q$ is a contraction if

$$\left(1+\frac{\alpha_{\max}}{\alpha_{\min}}\right)c(\Omega)L_\alpha\|\mathbf{u}\|_{\mathbf{L}^3} < 1,$$

which is condition (3.12). □

Remark 6. We see that (3.12) coincides with the condition for global uniqueness (3.7). This reflects that the nonlinearity G does not have regularizing properties. Nevertheless, these are only sufficient conditions, and it is plausible that problem (1.1) has a nonsingular solution without satisfying condition (3.12).

B. Discretization

Having analyzed the mathematical properties of problem (1.1) we now proceed to propose several methods for its approximate solution. With this purpose, let h be a discretization parameter (that will tend to zero). For every $h > 0$ we introduce two finite dimensional spaces $\mathbf{X}_h \subset \mathbf{L}^2(\Omega)$ and $M_h \subset H^1_w(\Omega)$ such that:

1. The pair of spaces $(\mathbf{X}_h, M_h)$ is stable, in the sense that they satisfy a uniform inf–sup condition (cf. [16, 34], A. Ern and J.-L. Guermond [27] or D. Boffi, et al. [13]). That is, there exists a constant $\beta > 0$ independent of h such that

$$\sup_{\mathbf{w}_h \in \mathbf{X}_h} \frac{b(\mathbf{w}_h, q_h)}{\|\mathbf{w}_h\|_{\mathbf{L}^2}} \geq \beta|q_h|_{H^1}, \quad \forall q_h \in M_h, \tag{3.17}$$

   where the form b is defined in (3.5).

2. There exist continuous interpolation operators $\pi_h : \mathbf{L}^2(\Omega) \to \mathbf{X}_h$, $\mathcal{I}_h : H^1(\Omega) \to M_h$ and an integer $\ell \geq 1$, such that for all $(\mathbf{v}, q) \in H^\ell(\Omega) \times H^{\ell+1}(\Omega)$

$$\|\mathbf{v} - \pi_h \mathbf{v}\|_{\mathbf{L}^2} \leq c h^\ell \|\mathbf{v}\|_{\mathbf{H}^\ell}, \tag{3.18}$$

and

$$|q - \mathcal{I}_h q|_{H^1} \leq c h^\ell |q|_{H^{\ell+1}}. \tag{3.19}$$

In order to find examples of such discrete spaces, assume to simplify that Ω is a polyhedron, and let $\mathcal{T}_h$ be a family of triangulations of $\bar{\Omega}$, made of tetrahedra with diameter bounded by h. We suppose that $\mathcal{T}_h$ is regular in the following sense (cf. P.G.Ciarlet [21]): There exists a constant $\sigma > 0$, independent of h, such that

$$\frac{h_T}{\rho_T} \leq \sigma, \quad \forall T \in \mathcal{T}_h, \tag{3.20}$$

where h_T is the diameter of T and ρ_T is the diameter of the ball inscribed in T. Then, for any integer $k \geq 1$, the following pair of spaces satisfy conditions (3.17)–(3.19):

$$\mathbf{X}_h := \left\{ \mathbf{v}_h \in \mathbf{L}^2(\Omega) : \ \mathbf{v}_h|_T \in \mathbb{P}_{k-1}^d, \forall T \in \mathcal{T}_h \right\}, \tag{3.21}$$

and

$$M_h := \left\{ q_h \in \mathcal{C}^0(\bar{\Omega}) : \ q_h|_T \in \mathbb{P}_k, \forall T \in \mathcal{T}_h \right\}. \tag{3.22}$$

For a proof the reader can consult standard references, for instance [16, 34, 27]. Finally, we define the discrete solution space

$$\mathfrak{X}_h := \mathbf{X}_h \times M_h,$$

normed by $\|\cdot\|_{\mathfrak{X}}$. Clearly, $\mathbf{X}_h \subset \mathfrak{X}$. For the sequel, it is also useful to introduce the space

$$\mathbf{V}_h := \{ \mathbf{v}_h \in \mathbf{X}_h : \ \forall q_h \in M_h \ b(\mathbf{v}_h, q_h) = 0 \}, \tag{3.23}$$

and its orthogonal in $\mathbf{X}_h$

$$\mathbf{V}_h^{\perp} = \{\mathbf{v}_h \in \mathbf{X}_h : \ \forall \mathbf{w}_h \in \mathbf{V}_h \ \int_\Omega \mathbf{v}_h \cdot \mathbf{w}_h = 0\}. \tag{3.24}$$

For each such pair of discrete spaces we define the Galerkin solution to problem (3.3) as the pair $x_h = (\mathbf{u}_h, p_h) \in \mathbf{X}_h$ such that

$$\begin{cases} a(p_h; \mathbf{u}_h, \mathbf{v}_h) + b(\mathbf{v}_h, p_h) = \int_\Omega \mathbf{f} \cdot \mathbf{v}_h, & \forall \mathbf{v}_h \in \mathbf{X}_h, \\ b(\mathbf{u}_h, q_h) = \langle g, q_h \rangle_\Gamma, & \forall q_h \in M_h. \end{cases} \tag{3.25}$$

Under assumptions (3.1) and (3.17), the existence of a solution for this problem can be established by the same techniques used in Theorem 1 (cf. [5]). It is even simpler, since problem (3.25) is already set in finite dimension. All solutions of problem (3.25) satisfy uniform a priori estimates and (3.18) and (3.19) suffice to establish weak convergence (up to subsequences) of any solution of (3.25) to some solution of (3.3).

In the remainder of this Section we analyze this discrete problem. For the case when the solution is unique we prove optimal error estimates and propose an algorithm to find such an approximate solution. The algorithm is proved to converge independently of the discretization parameter. For the nonuniqueness case, in the spirit of [17, 34],we show that for h small enough there exists a nonsingular solution to (3.25) in a neighborhood of the nonsingular solution to the exact problem. We analyze some properties of the application of Newton's method to this problem, and we obtain estimates on its speed of convergence and conditions on the initial approximation. The main difficulty in this analysis is that there exist x in $\mathfrak{X}$ for which the operator $G'(x)$ is not bounded in $\mathcal{L}(\mathfrak{X}, \mathfrak{Y})$. More precisely, we require that the first component of x belong to $\mathbf{L}^3(\Omega)$, a smaller space than $\mathbf{L}^2(\Omega)$. This again is related to the fact that the nonlinearity G does not have regularizing properties.

Recall that condition (3.7) is sufficient for the solution to problem (3.3) to be unique. In the setting that we have described, and under a similar assumption, we have the following *a priori* estimate.

Theorem 2. *Let the pair of finite dimensional spaces* $\mathbf{X}_h$ *satisfy condition* (3.17). *Assume that the solution* $x = (\mathbf{u}, p) \in \mathfrak{X}$ *to* (3.3) *is such that* $\mathbf{u} \in \mathbf{L}^3(\Omega)$ *and is small enough, in the sense that*

$$\frac{1}{\beta}\frac{\alpha_{\max} + \alpha_{\min}}{\alpha_{\min}} c(\Omega) L_\alpha \|\mathbf{u}\|_{\mathbf{L}^3} \leq \theta < 1. \tag{3.26}$$

Then both (3.3) *and* (3.25) *have a unique solution and there exists a constant* $c > 0$ *independent of* h *such that the solution* $x_h = (\mathbf{u}_h, p_h) \in \mathbf{X}_h$ *of problem* (3.25) *satisfies*

$$\|\mathbf{u} - \mathbf{u}_h\|_{\mathbf{L}^2} + |p - p_h|_{H^1} \leq c \left(\inf_{\mathbf{v}_h \in \mathbf{X}_h} \|\mathbf{u} - \mathbf{v}_h\|_{\mathbf{L}^2} + \inf_{q_h \in M_h} |p - q_h|_{H^1} \right). \tag{3.27}$$

Proof. The proof proceeds in three steps.

(i) The second equation in (3.25) can be viewed as a non-homogeneous constraint; let us show that we can approximate $\mathbf{u}$ with functions of $\mathbf{X}_h$ that satisfy this constraint. For this, let $\mathbf{v}_h$ be an arbitrary function of $\mathbf{X}_h$, define $\mathbf{r}_h$ in $\mathbf{X}_h$ by

$$b(\mathbf{r}_h, q_h) = b(\mathbf{u} - \mathbf{v}_h, q_h), \quad \forall q_h \in M_h,$$

and set $\mathbf{w}_h := \mathbf{r}_h + \mathbf{v}_h$. It follows from (3.17) and the Babuška–Brezzi's theory (cf. [6] or [16, 34, 27]) that this equation has a solution $\mathbf{r}_h \in \mathbf{X}_h$, unique in $\mathbf{V}_h^\perp$, and such that

$$\beta \|\mathbf{r}_h\|_{\mathbf{L}^2} \leq \|\mathbf{u} - \mathbf{v}_h\|_{\mathbf{L}^2}. \tag{3.28}$$

Thus

$$b(\mathbf{w}_h, q_h) = b(\mathbf{u}, q_h) = \langle g, q_h \rangle = b(\mathbf{u}_h, q_h), \quad \forall q_h \in M_h,$$

and $\mathbf{u}_h - \mathbf{w}_h \in \mathbf{V}_h$. This implies

$$\begin{aligned}\alpha_{\min}\|\mathbf{u}_h - \mathbf{w}_h\|_{\mathbf{L}^2} &\leq \sup_{0\neq\mathbf{y}_h\in\mathbf{V}_h} \frac{a(p_h;\mathbf{u}_h-\mathbf{w}_h,\mathbf{y}_h)}{\|\mathbf{y}_h\|_{\mathbf{L}^2}} \\ &\leq \sup_{0\neq\mathbf{y}_h\in\mathbf{V}_h} \frac{a(p_h;\mathbf{u}_h-\mathbf{u},\mathbf{y}_h)}{\|\mathbf{y}_h\|_{\mathbf{L}^2}} + \sup_{0\neq\mathbf{y}_h\in\mathbf{V}_h} \frac{a(p_h;\mathbf{u}-\mathbf{w}_h,\mathbf{y}_h)}{\|\mathbf{y}_h\|_{\mathbf{L}^2}} \\ &\leq \sup_{0\neq\mathbf{y}_h\in\mathbf{V}_h} \frac{a(p_h;\mathbf{u}_h-\mathbf{u},\mathbf{y}_h)}{\|\mathbf{y}_h\|_{\mathbf{L}^2}} + \alpha_{\max}\|\mathbf{u}-\mathbf{w}_h\|_{\mathbf{L}^2}.\end{aligned}$$

(ii) Subtract the first equation of (3.3) from the first equation in (3.25) with test function $\mathbf{y}_h \in \mathbf{V}_h$. Since $\mathbf{X}_h \subset \mathbf{L}^2(\Omega)$,

$$\begin{aligned}a(p_h;\mathbf{u}_h-\mathbf{u},\mathbf{y}_h) &= \int_\Omega (\alpha(p)-\alpha(p_h))\,\mathbf{u}\cdot\mathbf{y}_h + \int_\Omega \mathbf{y}_h\cdot\nabla(p-p_h) \\ &\leq L_\alpha\|p-p_h\|_{L^6}\|\mathbf{u}\|_{\mathbf{L}^3}\|\mathbf{y}_h\|_{\mathbf{L}^2} + b(\mathbf{y}_h,p-p_h) \\ &\leq c(\Omega)L_\alpha|p-p_h|_{H^1}\|\mathbf{u}\|_{\mathbf{L}^3}\|\mathbf{y}_h\|_{\mathbf{L}^2} + b(\mathbf{y}_h,p-q_h) + b(\mathbf{y}_h,q_h-p_h).\end{aligned}$$

This yields

$$\alpha_{\min}\|\mathbf{u}_h-\mathbf{w}_h\|_{\mathbf{L}^2} \leq c(\Omega)L_\alpha|p-p_h|_{H^1}\|\mathbf{u}\|_{\mathbf{L}^3} + |p-q_h|_{H^1} + \alpha_{\max}\|\mathbf{u}-\mathbf{w}_h\|_{\mathbf{L}^2},$$

where the last inequality holds since $\mathbf{y}_h \in \mathbf{V}_h$. Finally, by the triangle inequality and (3.28)

$$\begin{aligned}\alpha_{\min}\|\mathbf{u}-\mathbf{u}_h\|_{\mathbf{L}^2} \leq (\alpha_{\min}+\alpha_{\max})\left(1+\frac{1}{\beta}\right)\|\mathbf{u}-\mathbf{v}_h\|_{\mathbf{L}^2} \\ + c(\Omega)L_\alpha|p-p_h|_{H^1}\|\mathbf{u}\|_{\mathbf{L}^3} + |p-q_h|_{H^1}. \quad (3.29)\end{aligned}$$

(iii) Let $q_h \in M_h$ be arbitrary. By the inf–sup condition (3.17),

$$\begin{aligned}\beta|p_h - q_h|_{H^1} &\leq \sup_{0\neq \mathbf{y}_h\in\mathbf{X}_h} \frac{b(\mathbf{y}_h, p_h - q_h)}{\|\mathbf{y}_h\|_{\mathbf{L}^2}} \\ &\leq \sup_{0\neq \mathbf{y}_h\in\mathbf{X}_h} \frac{b(\mathbf{y}_h, p_h - p)}{\|\mathbf{y}_h\|_{\mathbf{L}^2}} + \sup_{0\neq \mathbf{y}_h\in\mathbf{X}_h} \frac{b(\mathbf{y}_h, p - q_h)}{\|\mathbf{y}_h\|_{\mathbf{L}^2}} \\ &\leq \sup_{0\neq \mathbf{y}_h\in\mathbf{X}_h} \frac{b(\mathbf{y}_h, p_h - p)}{\|\mathbf{y}_h\|_{\mathbf{L}^2}} + |p - q_h|_{H^1}.\end{aligned}$$

Subtracting the first equation of (3.3) from the first equation of (3.25), since $\mathbf{X}_h \subset \mathbf{L}^2(\Omega)$ we obtain

$$\begin{aligned}b(\mathbf{y}_h, p_h - p) &= \int_\Omega (\alpha(p) - \alpha(p_h))\, \mathbf{u}\cdot\mathbf{y}_h + \int_\Omega \alpha(p_h)(\mathbf{u} - \mathbf{u}_h)\cdot\mathbf{y}_h \\ &\leq c(\Omega)L_\alpha|p - p_h|_{H^1}\|\mathbf{u}\|_{\mathbf{L}^3}\|\mathbf{y}_h\|_{\mathbf{L}^2} + \alpha_{\max}\|\mathbf{u} - \mathbf{u}_h\|_{\mathbf{L}^2}\|\mathbf{y}_h\|_{\mathbf{L}^2},\end{aligned}$$

which implies

$$|p_h - q_h|_{H^1} \leq \frac{1}{\beta}|p - q_h|_{H^1} + \frac{c(\Omega)L_\alpha}{\beta}\|\mathbf{u}\|_{\mathbf{L}^3}|p - p_h|_{H^1} + \frac{\alpha_{\max}}{\beta}\|\mathbf{u} - \mathbf{u}_h\|_{\mathbf{L}^2}.$$

By the triangle inequality

$$|p - p_h|_{H^1} \leq \left(1 + \frac{1}{\beta}\right)|p - q_h|_{H^1} + \frac{c(\Omega)L_\alpha}{\beta}\|\mathbf{u}\|_{\mathbf{L}^3}|p - p_h|_{H^1} + \frac{\alpha_{\max}}{\beta}\|\mathbf{u} - \mathbf{u}_h\|_{\mathbf{L}^2}.$$

Assumption (3.26) implies

$$\frac{\alpha_{\max} + \alpha_{\min}(1-\theta)}{\alpha_{\max} + \alpha_{\min}}|p - p_h|_{H^1} \leq \left(1 + \frac{1}{\beta}\right)|p - q_h|_{H^1} + \frac{\alpha_{\max}}{\beta}\|\mathbf{u} - \mathbf{u}_h\|_{\mathbf{L}^2}.$$

Combining this last inequality, assumption (3.26), and (3.29) we obtain

$$\|\mathbf{u} - \mathbf{u}_h\|_{\mathbf{L}^2} \leq c\,(\|\mathbf{u} - \mathbf{v}_h\|_{\mathbf{L}^2} + |p - q_h|_{H^1}) + \frac{\alpha_{\max}\theta}{\alpha_{\max} + \alpha_{\min}(1-\theta)}\|\mathbf{u} - \mathbf{u}_h\|_{\mathbf{L}^2}.$$

Since

$$1 - \frac{\alpha_{\max}\theta}{\alpha_{\max} + \alpha_{\min}(1-\theta)} = \frac{\alpha_{\max} + \alpha_{\min}}{\alpha_{\max} + \alpha_{\min}(1-\theta)}(1-\theta) > 0$$

and the pair $(\mathbf{v}_h, q_h) \in \mathbf{X}_h$ is arbitrary we obtain the desired result. □

Remark 7. For the pair of finite element spaces (3.21), (3.22) condition (3.17) holds with $\beta = 1$. Hence, in this case, assumption (3.26) is the same as (3.7).

The next corollary follows readily from this Theorem.

Corollary 1. *Under the setting of Theorem 2, if the spaces* $\mathbf{X}_h$ *and* M_h *satisfy assumptions* (3.18) *and* (3.19)*, then*

$$\lim_{h\to 0} \|(\mathbf{u}, p) - (\mathbf{u}_h, p_h)\|_{\mathfrak{X}} = 0.$$

Moreover, if the exact solution $(\mathbf{u}, p) \in \mathbf{H}^s(\Omega) \times H^{s+1}(\Omega)$ *for some real number* $s \in [0, \ell]$*, then there is a constant* $c > 0$ *independent of* h *such that*

$$\|(\mathbf{u}, p) - (\mathbf{u}_h, p_h)\|_{\mathfrak{X}} \leq ch^s \left(\|\mathbf{u}\|_{\mathbf{H}^s} + \|q\|_{H^{s+1}}\right).$$

Proof. The conclusion of Theorem 2, an elementary density argument and assumptions (3.18) and (3.19) give that the Galerkin solution converges to the exact solution as $h \to 0$. If the exact solution is more regular, assumptions (3.18) and (3.19) give the claimed error estimates. □

We now propose an iterative scheme to solve the discrete nonlinear system (3.25). Although the scheme requires assembling a new matrix at each iterative step, we show that, under an assumption similar to (3.7), the speed of convergence to the Galerkin solution is independent of the discretization parameter h.

The proposed scheme is the following:

Given an arbitrary initial approximation $p_h^{(0)} \in M_h$, *for* $n \geq 0$ *find* $(\mathbf{u}_h^{(n+1)}, p_h^{(n+1)}) \in \mathbf{X}_h$ *that solve*

$$\begin{cases} a\left(p_h^{(n)}; \mathbf{u}_h^{(n+1)}, \mathbf{v}_h\right) + b\left(\mathbf{v}_h, p_h^{(n+1)}\right) = \int_\Omega \mathbf{f}\cdot\mathbf{v}_h, & \forall \mathbf{v}_h \in \mathbf{X}_h, \\ b\left(\mathbf{u}_h^{(n+1)}, q_h\right) = \langle g, q_h\rangle_\Gamma, & \forall q_h \in M_h. \end{cases} \tag{3.30}$$

Now we prove that this scheme converges independently of the discretization parameter.

Proposition 4. *Assume that the pair of spaces* $(\mathbf{X}_h, M_h)$ *satisfies condition* (3.17). *Let the solution to* (3.25) *be small enough, in the sense that there are two constants* $\theta < 1$ *and* $h_0 > 0$ *such that for every* $h \leq h_0$

$$\frac{\alpha_{\max} + \alpha_{\min}}{\alpha_{\min}} c(\Omega) L_\alpha \|\mathbf{u}_h\|_{\mathbf{L}^3} \leq \theta. \tag{3.31}$$

Then for the iterative scheme (3.30) *the following error estimates hold*

$$\left\|\mathbf{u}_h - \mathbf{u}_h^{(n+1)}\right\|_{\mathbf{L}^2} \leq \frac{1}{\alpha_{\max} + \alpha_{\min}} \frac{\theta^{n+1}}{\beta^n} \left|p_h - p_h^{(0)}\right|_{H^1},$$

and

$$\left|p_h - p_h^{(n+1)}\right|_{H^1} \leq \left(\frac{\theta}{\beta}\right)^{n+1} \left|p_h - p_h^{(0)}\right|_{H^1}.$$

Proof. Take the difference of equations (3.25) and (3.30). We obtain

$$\begin{cases} \int_\Omega \left(\alpha(p_h)\mathbf{u}_h - \alpha\left(p_h^{(n)}\right)\mathbf{u}_h^{(n+1)}\right)\cdot\mathbf{v}_h + b\left(\mathbf{v}_h, p_h - p_h^{(n+1)}\right) = 0, & \forall \mathbf{v}_h \in \mathbf{X}_h, \\ b\left(\mathbf{u}_h - \mathbf{u}_h^{(n+1)}, q_h\right) = 0, & \forall q_h \in M_h. \end{cases}$$

Set $\mathbf{v}_h = \mathbf{u}_h - \mathbf{u}_h^{(n+1)}$, then

$$\begin{aligned}\alpha_{\min}\left\|\mathbf{u}_h - \mathbf{u}_h^{(n+1)}\right\|_{\mathbf{L}^2}^2 &\leq \left|\int_\Omega \left(\alpha\left(p_h^{(n)}\right) - \alpha(p_h)\right)\mathbf{u}_h\cdot\left(\mathbf{u}_h - \mathbf{u}_h^{(n+1)}\right)\right| \\ &\leq c(\Omega)L_\alpha\left|p_h - p_h^{(n)}\right|_{H^1}\|\mathbf{u}_h\|_{\mathbf{L}^3}\left\|\mathbf{u}_h - \mathbf{u}_h^{(n+1)}\right\|_{\mathbf{L}^2},\end{aligned}$$

which by (3.31) implies

$$\left\|\mathbf{u}_h - \mathbf{u}_h^{(n+1)}\right\|_{\mathbf{L}^2} \leq \frac{\theta}{\alpha_{\max}+\alpha_{\min}}\left|p_h - p_h^{(n)}\right|_{H^1}. \tag{3.32}$$

By the inf–sup condition (3.17),

$$\begin{aligned}\beta\left|p_h - p_h^{(n+1)}\right|_{H^1} &\leq \sup_{0\neq\mathbf{v}_h\in\mathbf{X}_h}\frac{b\left(\mathbf{v}_h, p_h - p_h^{(n+1)}\right)}{\|\mathbf{v}_h\|_{\mathbf{L}^2}} \\ &= \sup_{0\neq\mathbf{v}_h\in\mathbf{X}_h}\frac{\int_\Omega\left(\alpha(p_h)\mathbf{u}_h - \alpha\left(p_h^{(n)}\right)\mathbf{u}_h^{(n+1)}\right)\cdot\mathbf{v}_h}{\|\mathbf{v}_h\|_{\mathbf{L}^2}} \\ &\leq \sup_{0\neq\mathbf{v}_h\in\mathbf{X}_h}\frac{\int_\Omega\left(\alpha(p_h) - \alpha(p_h^{(n)})\right)\mathbf{u}_h\cdot\mathbf{v}_h}{\|\mathbf{v}_h\|_{\mathbf{L}^2}} \\ &+ \sup_{0\neq\mathbf{v}_h\in\mathbf{X}_h}\frac{\int_\Omega\alpha\left(p_h^{(n)}\right)\left(\mathbf{u}_h - \mathbf{u}_h^{(n+1)}\right)\cdot\mathbf{v}_h}{\|\mathbf{v}_h\|_{\mathbf{L}^2}} \\ &\leq c(\Omega)L_\alpha\left|p_h - p_h^{(n)}\right|_{H^1}\|\mathbf{u}_h\|_{\mathbf{L}^3} + \alpha_{\max}\left\|\mathbf{u}_h - \mathbf{u}_h^{(n+1)}\right\|_{\mathbf{L}^2}.\end{aligned}$$

By condition (3.31) and inequality (3.32)

$$\begin{aligned}\beta\left|p_h - p_h^{(n+1)}\right|_{H^1} &\leq \frac{\theta}{\alpha_{\max}+\alpha_{\min}}\left(\alpha_{\max}+\alpha_{\min}\right)\left|p_h - p_h^{(n)}\right|_{H^1} \\ &= \theta\left|p_h - p_h^{(n)}\right|_{H^1}.\end{aligned}$$

From this inequality and (3.32) the claimed error bounds follow. □

Remark 8. One might argue that the previous error bounds do not guarantee convergence of the algorithm, since the value of β is not known and, hence, the ratio θ/β

could be greater than one. Using a similar assumption as (3.26), namely

$$\frac{1}{\beta}\frac{\alpha_{\max}+\alpha_{\min}}{\alpha_{\min}}c(\Omega)L_\alpha\|\mathbf{u}_h\|_{\mathbf{L}^3}\leq\theta,$$

we can bypass this constraint. Moreover, as we have mentioned before, for the concrete examples of spaces (3.21)–(3.22) we have $\beta=1$.

Remark 9. In addition to (3.17)–(3.19), assume that the following *inverse inequality* holds

$$\|\mathbf{v}_h\|_{\mathbf{L}^3}\leq ch^{-1/2}\|\mathbf{v}_h\|_{\mathbf{L}^2},\quad\forall\mathbf{v}_h\in\mathbf{X}_h. \tag{3.33}$$

If the exact solution $(\mathbf{u},p)$ belongs to $\mathbf{H}^s(\Omega)\times H^{s+1}(\Omega)$ for some real number s with $\frac{1}{2}<s\leq 1$, then the uniqueness condition (3.7) implies (3.31). Indeed, under these assumptions we have

$$\|\mathbf{u}-\mathbf{u}_h\|_{\mathbf{L}^3}=\mathcal{O}(h^{s-\frac{1}{2}}),$$

hence, if

$$\frac{\alpha_{\max}+\alpha_{\min}}{\alpha_{\min}}c(\Omega)L_\alpha\|\mathbf{u}\|_{\mathbf{L}^3}\leq\Theta<1,$$

then,

$$\frac{\alpha_{\max}+\alpha_{\min}}{\alpha_{\min}}c(\Omega)L_\alpha\|\mathbf{u}_h\|_{\mathbf{L}^3}\leq(1+\mathcal{O}(h^{s-\frac{1}{2}}))\Theta.$$

If h is small enough, we obtain condition (3.31).

Let us study now the approximation of nonsingular solutions. With this purpose, we introduce a final assumption on the function α, namely

$$\alpha\in W^2_\infty(\mathbb{R}). \tag{3.34}$$

As we have mentioned before, in the truncated case this is not restrictive for the problem we are treating.

Next, we complement (3.17)–(3.19) and (3.33) with an additional *inverse inequality*:

$$\|q_h\|_{L^\infty} \leq ch^{-1/2}|q_h|_{H^1}, \quad \forall q_h \in M_h. \tag{3.35}$$

Both inverse inequalities (3.33) and (3.35) hold when the family of triangulations $\mathcal{T}_h$ is quasi-uniform (or uniformly regular) in the following sense (cf. [21]): In addition to (3.20), there exists a constant $\tau > 0$, independent of h, such that

$$h_T \geq \tau h, \quad \forall T \in \mathcal{T}_h. \tag{3.36}$$

We are now concerned with the approximation of nonsingular solutions to (3.10) under the hypotheses (3.17)–(3.19), (3.33), and (3.35). In order to do that, let us define the discrete solution operator to the linear Darcy equations $T_h : \mathfrak{Y} \to \mathbf{X}_h$. That is, for any $\eta = (\mathbf{f}, g) \in \mathfrak{Y}$, $\mathbf{X}_h \ni x_h = (\mathbf{u}_h, p_h) = T_h\eta = T_h(\mathbf{f}, g)$ solves

$$\begin{cases} a(\mathbf{u}_h, \mathbf{v}_h) + b(\mathbf{v}_h, p_h) = \int_\Omega \mathbf{f}\cdot\mathbf{v}_h, & \forall \mathbf{v}_h \in \mathbf{X}_h, \\ b(\mathbf{u}_h, q_h) = \langle g, q_h\rangle_\Gamma, & \forall q_h \in M_h, \end{cases}$$

where the bilinear form $a : \mathbf{L}^2(\Omega) \times \mathbf{L}^2(\Omega)$ is defined by

$$a(\mathbf{u}, \mathbf{v}) := \bar{\alpha} \int_\Omega \mathbf{u}\cdot\mathbf{v}.$$

It is a classical matter ([16, 27]) to show that, under assumption (3.17), this operator is well-defined, injective, $T_h \in \mathcal{L}(\mathfrak{Y}, \mathbf{X}_h)$, and there is a constant c independent of h such that

$$\|T_h(\mathbf{f}, g)\|_{\mathfrak{X}} \leq c\|(\mathbf{f}, g)\|_{\mathfrak{Y}}, \quad \forall (\mathbf{f}, g) \in \mathfrak{Y}. \tag{3.37}$$

We can also define the discrete nonlinearity. This is an operator $G_h : \mathbf{X}_h \to \mathbf{X}_h \times H_{00}^{1/2}(\Gamma)' \subset \mathfrak{Y}$, such that if $x_h = (\mathbf{u}_h, p_h) \in \mathbf{X}_h$, then $G_h(x_h) := (\mathbf{F}_h, -g)$, where

$\mathbf{F}_h \in \mathbf{X}_h$ is the unique solution to

$$\int_\Omega \mathbf{F}_h \cdot \mathbf{v}_h = \int_\Omega \left[(\alpha(p_h) - \bar{\alpha})\, \mathbf{u}_h - \mathbf{f} \right] \cdot \mathbf{v}_h, \quad \forall \mathbf{v}_h \in \mathbf{X}_h.$$

Finally, define the operator $F_h : \mathbf{X}_h \to \mathbf{X}_h$ by

$$F_h(x_h) := x_h + T_h G_h(x_h).$$

With this notation, problem (3.25) can be equivalently rewritten as:

Find $x_h \in \mathbf{X}_h$ such that

$$F_h(x_h) = 0. \tag{3.38}$$

The approximation properties of the operator T_h are the following.

Proposition 5. *Assume that* (3.17)–(3.19) *hold. Let $(\mathbf{f}, g) \in \mathfrak{Y}$ be such that $T(\mathbf{f}, g) \in \mathbf{H}^s(\Omega) \times H^{1+s}(\Omega) \subset \mathfrak{X}$, for some $0 < s \leq \ell$. Then, there is a constant $c > 0$, independent of h such that*

$$\|(T - T_h)(\mathbf{f}, g)\|_{\mathfrak{X}} \leq c h^s \|T(\mathbf{f}, g)\|_{\mathbf{H}^s \times H^{1+s}}. \tag{3.39}$$

Proof. It is a direct consequence of assumptions (3.17)–(3.19), together with a basic interpolation argument ([12]). □

Corollary 2. *Under the hypotheses of Proposition 5, the operator T_h satisfies*

$$\lim_{h \to 0} \|T - T_h\|_{\mathcal{L}(\mathfrak{Y}, \mathfrak{X})} = 0. \tag{3.40}$$

Proof. Standard regularity results for the linear Darcy problem (3.8) imply that, for sufficiently small $s > 0$, $T(\mathbf{f}, g) \in \mathbf{H}^s(\Omega) \times H^{1+s}(\Omega)$ if $(\mathbf{f}, g)$ belongs to $\tilde{\mathfrak{Y}} := \mathbf{H}^s(\Omega) \times H^{s-1/2}(\partial\Omega)$, which is a dense subset of $\mathfrak{Y}$. The boundedness of operator T

(see (3.9)), together with inequality (3.39) imply

$$\sup_{0\neq(\mathbf{f},g)\in\mathfrak{Y}}\frac{\|(T-T_h)(\mathbf{f},g)\|_{\mathfrak{X}}}{\|(\mathbf{f},g)\|_{\mathfrak{Y}}}=\sup_{(\mathbf{f},g)\in\tilde{\mathfrak{Y}}}\frac{\|(T-T_h)(\mathbf{f},g)\|_{\mathfrak{X}}}{\|(\mathbf{f},g)\|_{\mathfrak{Y}}}\leq ch^s\frac{\|T(\mathbf{f},g)\|_{\mathfrak{X}}}{\|(\mathbf{f},g)\|_{\mathfrak{Y}}}\leq ch^s,$$

from which (3.40) clearly follows. □

We are interested in approximating a nonsingular solution $x=(\mathbf{u},p)\in\mathfrak{X}$ to (3.10). For this, we must assume that there is a real number $s>1/2$ such that

$$(\mathbf{u},p)\in\mathbf{H}^s(\Omega)\times H^{1+s}(\Omega). \tag{3.41}$$

Remark 10. Since $s>1/2$, (3.41) implies that $(\mathbf{u},p)\in\mathbf{L}^3(\Omega)\times\mathcal{C}^0(\bar{\Omega})$, see [2].

To alleviate the notation, define

$$x_h^0:=(\mathbf{u}_h^0,p_h^0)=(\pi_h\mathbf{u},\mathcal{I}_hp)\in\mathbf{X}_h, \tag{3.42}$$

where π_h and $\mathcal{I}_h$ are the interpolation operators of (3.18) and (3.19) respectively. Important properties of the interpolant x_h^0 and the operator $F_h'(x_h^0)$ are established below.

Lemma 1. *Let the function α satisfy conditions* (3.1), (3.2) *and* (3.34). *Let the solution $(\mathbf{u},p)\in\mathfrak{X}$ to problem* (3.10) *be nonsingular and satisfy the smoothness condition* (3.41). *If the pair of spaces $(\mathbf{X}_h,M_h)$ satisfies assumptions* (3.18), (3.19), *then there exists a constant $c>0$ independent of h, such that*

$$\left\|\mathbf{u}-\mathbf{u}_h^0\right\|_{\mathbf{L}^2}\leq ch^s\|\mathbf{u}\|_{\mathbf{H}^s}, \tag{3.43}$$

and

$$\left|p-p_h^0\right|_{H^1}\leq ch^s\|p\|_{H^{1+s}}. \tag{3.44}$$

Moreover, if the pair $(\mathbf{X}_h,M_h)$ also satisfies conditions (3.17), (3.33) *and* (3.35), *then there exists a $h_0>0$ such that for every $h\leq h_0$ the operator $F_h'(x_h^0)$ is an isomorphism*

of $\mathbf{X}_h$ *and the norm of its inverse is bounded independently of* h.

Proof. Inequalities (3.43) and (3.44) are a simple consequence of (3.18), (3.19) and assumption (3.41) via interpolation ([12]).

To show that $F_h'(x_h^0)$ is an isomorphism of $\mathbf{X}_h$, notice that

$$I + T_h G_h'(x_h^0) = I + T_h G'(x) + T_h \left(G'(x_h^0) - G'(x)\right) + T_h \left(G_h'(x_h^0) - G'(x_h^0)\right).$$

Let us consider each term separately.

(i) $I + T_h G'(x)$. Notice, first of all, that if $y_h \in \mathbf{X}_h$, then $\left(I + T_h G'(x)\right) y_h \in \mathbf{X}_h$. Moreover,

$$I + T_h G'(x) - F'(x) = \left(T_h - T\right) G'(x).$$

Since x is a nonsingular solution, $F'(x)$ is an isomorphism of $\mathfrak{X}$. Corollary 2 and an application of the Theorem about the Perturbation of an Invertible Operator (see [58, Theorem 4, p.207] for instance) imply that there is $h_0^{(1)} > 0$ such that for all $h \leq h_0^{(1)}$ the operator $I + T_h G'(x)$ is an isomorphism of $\mathfrak{X}$. Hence it is an isomorphism of $\mathbf{X}_h$. Thus, the result of the Lemma will be proved if we show that the remaining two terms tend to zero (in the $\|\cdot\|_{\mathcal{L}(\mathbf{X}_h)}$–norm) as $h \to 0$.

(ii) $T_h(G'(x_h^0) - G'(x))$. Let $y_h = (\mathbf{v}_h, q_h)$; using the definition of the derivatives, for any $\mathbf{w} \in \mathbf{L}^2(\Omega)$

$$\begin{aligned}
\left\langle (G'(x_h^0) - G'(x)) y_h, (\mathbf{w}, 0) \right\rangle &= \int_\Omega \left(\alpha(p_h^0) - \alpha(p)\right) \mathbf{v}_h \cdot \mathbf{w} \\
&\quad + \int_\Omega \left(\dot\alpha(p_h^0)\mathbf{u}_h^0 - \dot\alpha(p)\mathbf{u}\right) q_h \cdot \mathbf{w} \\
&= \int_\Omega \left(\alpha(p_h^0) - \alpha(p)\right) \mathbf{v}_h \cdot \mathbf{w} \\
&\quad + \int_\Omega \left(\dot\alpha(p_h^0) - \dot\alpha(p)\right) q_h \mathbf{u} \cdot \mathbf{w} \\
&\quad + \int_\Omega \dot\alpha(p_h^0) \left(\mathbf{u}_h^0 - \mathbf{u}\right) q_h \cdot \mathbf{w}.
\end{aligned}$$

Consider each term separately. By (3.2) and the inverse inequality (3.33)

$$\begin{aligned}\int_\Omega \left(\alpha(p_h^0)-\alpha(p)\right)\mathbf{v}_h\cdot\mathbf{w} &\le c(\Omega)L_\alpha \left|p-p_h^0\right|_{H^1}\|\mathbf{v}_h\|_{\mathbf{L}^3}\|\mathbf{w}\|_{\mathbf{L}^2}\\ &\le ch^{-1/2}\left|p-p_h^0\right|_{H^1}\|\mathbf{v}_h\|_{\mathbf{L}^2}\|\mathbf{w}\|_{\mathbf{L}^2}.\end{aligned}$$

By (3.34) and the inverse inequality (3.35)

$$\begin{aligned}\int_\Omega \left(\dot\alpha(p_h^0)-\dot\alpha(p)\right)q_h\mathbf{u}\cdot\mathbf{w} &\le c\|q_h\|_{L^\infty}\int_\Omega \left|p-p_h^0\right||\mathbf{u}||\mathbf{w}|\\ &\le ch^{-1/2}\left|p-p_h^0\right|_{H^1}|q_h|_{H^1}\|\mathbf{u}\|_{\mathbf{L}^3}\|\mathbf{w}\|_{\mathbf{L}^2}.\end{aligned}$$

Finally, by (3.2) and the inverse inequality (3.35)

$$\begin{aligned}\int_\Omega \dot\alpha(p_h^0)\left(\mathbf{u}_h^0-\mathbf{u}\right)q_h\cdot\mathbf{w} &\le L_\alpha\left\|\mathbf{u}-\mathbf{u}_h^0\right\|_{\mathbf{L}^2}\|q_h\|_{L^\infty}\|\mathbf{w}\|_{\mathbf{L}^2}\\ &\le ch^{-1/2}\left\|\mathbf{u}-\mathbf{u}_h^0\right\|_{\mathbf{L}^2}|q_h|_{H^1}\|\mathbf{w}\|_{\mathbf{L}^2}.\end{aligned}$$

Thus, by the stability property (3.37) of T_h,

$$\begin{aligned}\left\|T_h(G'(x_h^0)-G'(x))\right\|_{\mathcal{L}(\mathbf{X}_h)} &\le c\|G'(x_h^0)-G'(x)\|_{\mathcal{L}(\mathbf{X}_h,\mathfrak{Y})}\\ &= c\sup_{0\neq y_h\in\mathbf{X}_h}\frac{\|(G'(x_h^0)-G'(x))y_h\|_{\mathfrak{Y}}}{\|y_h\|_{\mathfrak{X}}}\\ &= c\sup_{0\neq y_h\in\mathbf{X}_h}\sup_{0\neq\mathbf{w}\in\mathbf{L}^2(\Omega)}\frac{\langle(G'(x_h^0)-G'(x))y_h,\mathbf{w}\rangle}{\|y_h\|_{\mathfrak{X}}\|\mathbf{w}\|_{\mathbf{L}^2}}\\ &\le ch^{-1/2}\left(\left|p-p_h^0\right|_{H^1}+\left\|\mathbf{u}-\mathbf{u}_h^0\right\|_{\mathbf{L}^2}\right),\end{aligned}$$

which by the approximation properties (3.43) and (3.44) of x_h^0 and the fact that $s>1/2$ implies that this last quantity tends to zero as $h\to 0$.

(iii) $T_h(G_h'(x_h^0)-G'(x_h^0))$. It is sufficient to notice that for any $\mathbf{w}_h\in\mathbf{X}_h$

$$\left\langle(G_h'(x_h^0)-G'(x_h^0))y_h,(\mathbf{w}_h,0)\right\rangle=0.$$

□

Remark 11. In the example (3.21), (3.22), as in most finite element spaces, inverse estimates such as (3.33) and (3.35) hold locally. Therefore they may be applied locally when used in proving the interpolation Lemma 1, because interpolation properties are also local. In this case, the statement of Lemma 1 is valid even if the triangulation is not quasi-uniform. But of course intermediate results would have to be stated differently. For instance the bound for

$$\int_\Omega \big(\alpha(p_h^0) - \alpha(p)\big)\, \mathbf{v}_h \cdot \mathbf{w}$$

would read, for $s > \frac{1}{2}$:

$$\int_\Omega \big(\alpha(p_h^0) - \alpha(p)\big)\, \mathbf{v}_h \cdot \mathbf{w} \leq c h^{s-1/2} |p|_{H^{1+s}} \|\mathbf{v}_h\|_{\mathbf{L}^2} \|\mathbf{w}\|_{\mathbf{L}^2}.$$

However, this does not apply to inverse inequalities that are used in conjunction with global error estimates, such as in Remark 9 or in Lemma 2 below, in which case some restriction on the mesh cannot be avoided.

Once we know the main properties of the operator $F_h'(x_h^0)$, it is possible to study $F_h'(y_h)$ for y_h close to x_h^0.

Lemma 2. *Under the assumptions of Lemma 1, there is a constant $c_0 > 0$ independent of h such that*

$$\|G_h'(y_h) - G_h'(x_h^0)\|_{\mathcal{L}(\mathbf{X}_h, \mathfrak{Y})} \leq c_0 h^{-1/2} \left\|y_h - x_h^0\right\|_{\mathfrak{X}}, \quad \forall y_h \in \mathbf{X}_h. \tag{3.45}$$

Proof. Let $y_h = (\mathbf{v}_h, q_h), z_h = (\mathbf{w}_h, r_h) \in \mathbf{X}_h$. For an arbitrary $\mathbf{t}_h \in \mathbf{X}_h$

$$\begin{aligned}\left\langle (G_h'(y_h) - G_h'(x_h^0))z_h, (\mathbf{t}_h, 0)\right\rangle &= \int_\Omega \left(\alpha(q_h) - \alpha(p_h^0)\right) \mathbf{w}_h \cdot \mathbf{t}_h + \int_\Omega \dot\alpha(q_h) \left(\mathbf{v}_h - \mathbf{u}_h^0\right) r_h \cdot \mathbf{t}_h \\ &+ \int_\Omega \left(\dot\alpha(q_h) - \dot\alpha(p_h^0)\right) \mathbf{u}_h^0 r_h \cdot \mathbf{t}_h \\ &\le c \left(\left\| p_h^0 - q_h \right\|_{L^\infty} \|\mathbf{w}_h\|_{\mathbf{L}^2} \|\mathbf{t}_h\|_{\mathbf{L}^2} \right. \\ &+ \left\| \mathbf{u}_h^0 - \mathbf{v}_h \right\|_{\mathbf{L}^3} |r_h|_{H^1} \|\mathbf{t}_h\|_{\mathbf{L}^2} \\ &+ \left. \left\| p_h^0 - q_h \right\|_{L^\infty} \|\mathbf{u}_h^0\|_{\mathbf{L}^3} |r_h|_{H^1} \|\mathbf{t}_h\|_{\mathbf{L}^2} \right),\end{aligned}$$

hence

$$\|G_h'(y_h) - G_h'(x_h^0)\|_{\mathcal{L}(\mathbf{X}_h, \mathfrak{Y})} \le c \left(\left\| p_h^0 - q_h \right\|_{L^\infty} + \left\| \mathbf{u}_h^0 - \mathbf{v}_h \right\|_{\mathbf{L}^3} \right).$$

This estimate and the inverse inequalities (3.33), (3.35) imply (3.45). □

Remark 12. Lemma 2 states that G_h' is Lipschitz-continuous in a neighborhood of x_h^0, but this continuity is not uniform with respect to h. One more time, the absence of regularizing properties for the nonlinearity G does not allow us to obtain uniform in h bounds.

It is important to know whether the consistency error $F_h(x_h^0)$ tends to zero as $h \to 0$, and if this is the case at which rate. The following Lemma shows that the convergence is optimal given the regularity of the exact nonsingular solution x.

Lemma 3. *Under the assumptions of the first part of Lemma 1, there is a constant $c > 0$, independent of h such that*

$$\|F_h(x_h^0)\|_{\mathfrak{X}} \le c h^s \left(\|\mathbf{u}\|_{\mathbf{H}^s} + \|p\|_{H^{1+s}} \right). \tag{3.46}$$

Proof. Since $F(x) = 0$,

$$F_h(x_h^0) = x_h^0 - x + T_h(G_h(x_h^0) - G(x)) + (T_h - T)G(x),$$

which implies

$$\|F_h(x_h^0)\|_{\mathfrak{X}} \leq \|x - x_h^0\|_{\mathfrak{X}} + \|(T - T_h)G(x)\|_{\mathfrak{X}} + \|T_h(G(x) - G_h(x_h^0))\|_{\mathfrak{X}}.$$

From (3.43) and (3.44),

$$\|x - x_h^0\|_{\mathfrak{X}} \leq ch^s \left(\|\mathbf{u}\|_{\mathbf{H}^s} + \|p\|_{H^{1+s}}\right).$$

Estimate (3.39) implies

$$\|(T - T_h)G(x)\|_{\mathfrak{X}} \leq ch^s \|TG(x)\|_{\mathbf{H}^s \times H^{1+s}} = ch^s \left(\|\mathbf{u}\|_{\mathbf{H}^s} + \|p\|_{H^{1+s}}\right).$$

Finally, since $T_h(G_h(x_h^0) - G(x))$ belongs to $\mathbf{X}_h$, by the stability property (3.37) of T_h we see that it is sufficient to control the difference of the first coordinate of $G(x) - G_h(x_h^0)$ when tested against an element of $\mathbf{X}_h$. Let $\mathbf{v}_h \in \mathbf{X}_h$, then using (3.43) and (3.44)

$$\begin{aligned}\int_\Omega \left[G(x) - G_h(x_h^0)\right]_1 \cdot \mathbf{v}_h &\leq (\bar{\alpha} + \alpha_{\max})\|\mathbf{u} - \mathbf{u}_h^0\|_{\mathbf{L}^2}\|\mathbf{v}_h\|_{\mathbf{L}^2} \\ &\quad + c(\Omega)L_\alpha \left|p - p_h^0\right|_{H^1} \|\mathbf{u}\|_{\mathbf{L}^3}\|\mathbf{v}_h\|_{\mathbf{L}^2} \\ &\leq ch^s \left(\|\mathbf{u}\|_{\mathbf{H}^s} + \|p\|_{H^{1+s}}\right) \|\mathbf{v}_h\|_{\mathbf{L}^2}.\end{aligned}$$

□

According to the theory in [17, 34], Lemmas 1, 2, and 3 allow us to prove our main result, namely, the existence of a nonsingular solution for the discrete problem and optimal error estimates for it.

Theorem 3. *Let α satisfy* (3.1), (3.2) *and* (3.34). *Assume that problem* (3.10) *has a nonsingular solution $x = (\mathbf{u}, p) \in \mathbf{H}^s(\Omega) \times H^{1+s}(\Omega) \subset \mathfrak{X}$, for some $s > 1/2$. If the pair of spaces $(\mathbf{X}_h, M_h)$ satisfies* (3.17), (3.18), (3.19), (3.33), *and* (3.35), *then there is a $h_0 > 0$ such that for all $h \leq h_0$ the discrete problem* (3.38) *has a unique*

nonsingular solution $x_h = (\mathbf{u}_h, p_h)$ *in a neighborhood of the interpolant* $x_h^0 = (\mathbf{u}_h^0, p_h^0)$ *of the exact nonsingular solution. Moreover, this solution satisfies the following error estimate*

$$\|x - x_h\|_{\mathfrak{X}} \leq ch^s \left(\|\mathbf{u}\|_{\mathbf{H}^s} + \|p\|_{H^{1+s}} \right), \tag{3.47}$$

where the constant $c > 0$ *does not depend on* h.

Proof. Let us define

$$\epsilon_h := \|F_h(x_h^0)\|_{\mathfrak{X}},$$

and

$$\mathcal{M}_h(\delta) := \sup_{y_h \in \mathbf{X}_h:\ \|y_h - x_h^0\|_{\mathfrak{X}} < \delta} \|F_h'(y_h) - F_h'(x_h^0)\|_{\mathfrak{X}}.$$

Lemma 1 implies that there is a $h_0^{(1)} > 0$ such that for all $h \leq h_0^{(1)}$ the operator $F_h'(x_h^0)$ is an isomorphism of $\mathbf{X}_h$ with inverse bounded independently of h. Denote this bound by Δ. Inequalities (3.45) and (3.46) imply that

$$2\Delta \mathcal{M}_h(2\Delta\epsilon_h) \leq ch^{s-1/2},$$

hence there is a $h_0^{(2)} > 0$ such that for all $h \leq h_0^{(2)}$

$$2\Delta \mathcal{M}_h(2\Delta\epsilon_h) < 1.$$

Set $h_0 = \min\{h_0^{(1)}, h_0^{(2)}\}$ and consider $h \leq h_0$.

Since the operator $F_h'(x_h^0)$ is an isomorphism, solving problem (3.38) is equivalent to finding a fixed point of the map $\Phi_h : \mathbf{X}_h \to \mathbf{X}_h$ defined by

$$\Phi_h(y_h) := y_h - \left[F_h'(x_h^0)\right]^{-1} F_h(y_h).$$

Denote

$$S := \left\{ y_h \in \mathbf{X}_h :\ \|y_h - x_h^0\|_{\mathfrak{X}} \leq 2\Delta\epsilon_h \right\}.$$

We shall show that Φ_h is a contraction from S to S.

If $y_h \in S$,

$$\Phi_h(y_h) - x_h^0 = \left[F_h'(x_h^0)\right]^{-1} \left(F_h'(x_h^0)(y_h - x_h^0) - \left(F_h(y_h) - F_h(x_h^0)\right) - F_h(x_h^0)\right).$$

By the Mean Value Theorem

$$F_h(y_h) - F_h(x_h^0) = \int_0^1 F_h'\left(x_h^0 + \theta(y_h - x_h^0)\right)(y_h - x_h^0)\mathrm{d}\theta,$$

from which follows

$$\begin{aligned}
&\|F_h'(x_h^0)(y_h - x_h^0) - \left(F_h(y_h) - F_h(x_h^0)\right)\|_{\mathfrak{X}} \\
&\quad \leq \int_0^1 \left\|F_h'(x_h^0) - F_h'(x_h^0 + \theta(y_h - x_h^0))\right\|_{\mathcal{L}(\mathbf{X}_h)} \|y_h - x_h^0\|_{\mathfrak{X}}\mathrm{d}\theta \leq 2\Delta\epsilon_h\mathcal{M}_h(2\Delta\epsilon_h).
\end{aligned}$$

And, by the choice of h

$$\|\Phi_h(y_h) - x_h^0\|_{\mathfrak{X}} \leq \Delta\left(2\Delta\epsilon_h\mathcal{M}_h(2\Delta\epsilon_h) + \epsilon_h\right) = \Delta\epsilon_h\left(2\Delta\mathcal{M}_h(2\Delta\epsilon_h) + 1\right) < 2\Delta\epsilon_h,$$

which means that $\Phi_h(y_h) \in S$.

Let $y_h, z_h \in S$, then a similar computation shows that

$$\|\Phi_h(y_h) - \Phi_h(z_h)\|_{\mathfrak{X}} \leq \Delta\mathcal{M}_h(2\Delta\epsilon_h)\|y_h - z_h\|_{\mathfrak{X}} < \frac{1}{2}\|y_h - z_h\|_{\mathfrak{X}},$$

which implies that Φ_h is a contraction and we can conclude that there is a unique $x_h \in S$ such that $x_h = \Phi_h(x_h)$.

To realize that this solution is nonsingular, notice that

$$\left\|F_h'(x_h^0) - F_h'(x_h)\right\|_{\mathcal{L}(\mathbf{X}_h)} \leq \mathcal{M}_h(2\Delta\epsilon_h) < \frac{1}{2\Delta},$$

and apply the Theorem about the Perturbation of an Invertible Operator (see L.V. Kantorovich and G.P. Akilov [58, Theorem 4, p.207] for instance).

Finally, to get the error estimate (3.47) it is sufficient to use (3.46), the triangle inequality; and properties (3.43) and (3.44) of x_h^0,

$$\begin{aligned}\|x_h - x\|_{\mathfrak{X}} &\leq \|x_h - x_h^0\|_{\mathfrak{X}} + \|x_h^0 - x\|_{\mathfrak{X}} \\ &\leq 2\Delta\epsilon_h + ch^s \left(\|\mathbf{u}\|_{\mathbf{H}^s} + \|p\|_{H^{1+s}}\right) \\ &\leq ch^s \left(\|\mathbf{u}\|_{\mathbf{H}^s} + \|p\|_{H^{1+s}}\right).\end{aligned}$$

This concludes the proof. □

Remark 13. From the proof of this Theorem we see that the discrete nonsingular solution x_h is unique in a ball larger than S. Namely, it is unique in the ball

$$S(\bar{\delta}) := \left\{y_h \in \mathbf{X}_h : \ \|y_h - x_h^0\|_{\mathfrak{X}} < \bar{\delta}\right\},$$

where $\bar{\delta}$ is such that $\Delta\mathcal{M}_h(\bar{\delta}) < 1$. Both radii tend to zero as $h \to 0$. But, according to (3.46), the radius of S is $\mathcal{O}(h^s)$, $s > 1/2$, whereas $\bar{\delta} = \mathcal{O}(h^{1/2})$.

We have obtained that the discrete problem (3.38) has a unique nonsingular solution in a neighborhood of the exact nonsingular solution. We now analyze the application of Newton's method to the solution of this discrete problem. The algorithm is the following:

Given $x_h^{(0)} \in \mathbf{X}_h$, *for* $n \geq 0$ *define* $x_h^{(n+1)}$ *by*

$$x_h^{(n+1)} = x_h^{(n)} - \left[F_h'\left(x_h^{(n)}\right)\right]^{-1} F_h\left(x_h^{(n)}\right).$$

For this method to make sense $F_h'\left(x_h^{(n)}\right)$ must be an isomorphism of $\mathbf{X}_h$ for all n. Let us introduce the following notation

$$S(x_h, \delta) := \left\{y_h \in \mathbf{X}_h : \ \|y_h - x_h\|_{\mathfrak{X}} < \delta\right\},$$

and,

$$K := \frac{1}{4\|T_h\|_{\mathcal{L}(\mathfrak{Y},\mathbf{X}_h)} c_0 \Delta},$$

where the constant c_0 is the constant in inequality (3.45), Δ is such that for h small enough

$$\left\| \left[F_h'(x_h^0)\right]^{-1} \right\|_{\mathcal{L}(\mathbf{X}_h)} \leq \Delta,$$

and x_h^0 is the interpolant of x defined in (3.42).

Lemma 4. *There exists a real number $h_0 > 0$ such that for all $h \leq h_0$, if $\delta = \mathcal{O}(h^{1/2})$ and $y_h \in S(x_h, \delta)$, then the linear operator $F_h'(y_h)$ is an isomorphism of $\mathbf{X}_h$. Moreover, the norm of the inverse of this operator is bounded independently of h.*

Proof. Since

$$F_h'(y_h) = F_h'(x_h) + \left(F_h'(y_h) - F_h'(x_h)\right),$$

and, by Theorem 3, there exists $h_0 > 0$ such that for all $h \leq h_0$, $F_h'(x_h)$ is an isomorphism of $\mathbf{X}_h$, the result is obtained if we show that $F_h'(y_h) - F_h'(x_h)$ is small enough. We know that,

$$\left\| \left[F_h'(x_h)\right]^{-1} \right\|_{\mathcal{L}(\mathbf{X}_h)} \leq 2\Delta.$$

A similar argument as in the proof of Lemma 2 gives us that

$$\|F_h'(y_h) - F_h'(x_h)\|_{\mathcal{L}(\mathbf{X}_h)} \leq c_0 h^{-1/2} \|T_h\|_{\mathcal{L}(\mathfrak{Y},\mathbf{X}_h)} \|y_h - x_h\|_{\mathfrak{X}}.$$

Hence, if

$$2c_0 \|T_h\|_{\mathcal{L}(\mathfrak{Y},\mathbf{X}_h)} \|y_h - x_h\|_{\mathfrak{X}} \Delta h^{-1/2} < 1,$$

then the Theorem about the Perturbation of an Invertible Operator implies that $F_h'(y_h)$ is an isomorphism of $\mathbf{X}_h$. Moreover, from this inequality we see that it is sufficient to set

$$\delta \leq K h^{1/2},$$

where K is a constant independent of h. □

Theorem 4. *There exists a real number $h_0 > 0$ such that for all $h \leq h_0$, if*

$$\delta \leq \epsilon K h^{1/2},$$

for some real number ϵ with $0 < \epsilon < 1$, and if the initial approximation of Newton's method $x_h^{(0)}$ belongs to $S(x_h, \delta)$, then Newton's method converges to the discrete nonsingular solution x_h and the following error estimate holds

$$\left\| x_h^{(n+1)} - x_h \right\|_{\mathfrak{X}} \leq \frac{1}{K} h^{-1/2} \left\| x_h^{(n)} - x_h \right\|_{\mathfrak{X}}^2.$$

Proof. Assume h is small enough. Let us show by induction that if $x_h^{(0)} \in S(x_h, \delta)$, then $x_h^{(n)} \in S(x_h, \delta)$ for all $n > 0$. If $x_h^{(n)}$ is in $S(x_h, \delta)$ and δ is chosen as indicated, then by the previous Lemma, K can be chosen independently of h, so that $F_h'(x_h^{(n)})$ is an isomorphism of $\mathbf{X}_h$, with

$$\left\| \left[F_h' \left(x_h^{(n)} \right) \right]^{-1} \right\|_{\mathcal{L}(\mathbf{X}_h)} \leq 4\Delta.$$

Furthermore with a similar argument as in the proof of Theorem 3 we obtain

$$\begin{aligned} x_h^{(n+1)} - x_h &= \left[F_h' \left(x_h^{(n)} \right) \right]^{-1} \left(F_h' \left(x_h^{(n)} \right) \left(x_h^{(n)} - x_h \right) - \left(F_h \left(x_h^{(n)} \right) - F_h(x_h) \right) \right) \\ &= \left[F_h' \left(x_h^{(n)} \right) \right]^{-1} \int_0^1 \left[F_h' \left(x_h^{(n)} \right) - F_h' \left(x_h^{(n)} - \theta \left(x_h^{(n)} - x_h \right) \right) \right] \left(x_h^{(n)} - x_h \right) \mathrm{d}\theta. \end{aligned}$$

Then, by the induction hypothesis, a similar argument as in Lemma 2 and the choice

of δ and K imply

$$\begin{aligned}\left\|x_h^{(n+1)} - x_h\right\|_{\mathfrak{X}} &\leq \left\|\left[F_h'\left(x_h^{(n)}\right)\right]^{-1}\right\|_{\mathcal{L}(\mathbf{X}_h)} \|T_h\|_{\mathcal{L}(\mathfrak{Y},\mathbf{X}_h)} \times \\ &\quad \times \int_0^1 \left\|G_h'\left(x_h^{(n)}\right) - G_h'\left(x_h^{(n)} - \theta\left(x_h^{(n)} - x_h\right)\right)\right\|_{\mathcal{L}(\mathfrak{Y},\mathbf{X}_h)} \mathrm{d}\theta \left\|x_h^{(n)} - x_h\right\|_{\mathfrak{X}} \\ &\leq 4\Delta \|T_h\|_{\mathcal{L}(\mathfrak{Y},\mathbf{X}_h)} c_0 h^{-1/2} \left\|x_h^{(n)} - x_h\right\|_{\mathfrak{X}}^2 \\ &\leq \epsilon \left\|x_h^{(n)} - x_h\right\|_{\mathfrak{X}}.\end{aligned}$$

On one hand, this shows that $x_h^{(n+1)} \in S(x_h, \delta)$ and hence, by Lemma 4, that $F_h'(x_h^{(n+1)})$ is an isomorphism of $\mathbf{X}_h$ for all $n \geq 1$, on the other hand this shows the claimed error estimate. □

Remark 14. As we can see, the initial guess in Newton's method must be very close to the discrete solution. Moreover, the convergence of the method deteriorates as the discretization parameter h tends to zero. This is again related to the lack of regularizing properties for the nonlinearity G, as is reflected by Lemma 2.

C. A Splitting Algorithm for Exponential Porosity

The preceding analysis does not apply to an exponential porosity α, since assumptions (3.1) and (3.2) are not satisfied. So far, a rigorous analysis of this problem is beyond our reach. Nevertheless, for the exponential case, we propose a split formulation derived heuristically by taking the divergence of the first equation of (1.1) and making a change of variable.

Thus, by precisely exploiting the exponential character of the porosity (1.2), we are able to decompose the nonlinear Darcy problem into a linear elliptic equation and a linear Darcy system. But this process is heuristic since we develop this method without even knowing whether in general problem (1.1), with the porosity defined as

(1.2), does have a solution.

This section is organized as follows. First, we present the motivation behind the split formulation, next we study the properties of the solution to the auxiliary problem, i.e., the linear elliptic equation. Finally, we discretize the split formulation and we study the convergence of the resulting algorithm.

Let $(\mathbf{u}, p)$ be a solution of problem (1.1) with the porosity given by (1.2) and assume that p belongs to $L^\infty(\Omega)$. Since $\alpha(p) > 0$, we can divide the first equation in (1.1) by $\alpha(p)$, take the divergence of the result, and make a suitable change in variable. Using the second equation of (1.1), we obtain

$$0 = \nabla\cdot\mathbf{u} = \nabla\cdot\left(\frac{1}{\alpha(p)}\mathbf{f} - \frac{1}{\alpha(p)}\nabla p\right).$$

Since $1/\alpha(p) = 1/\alpha_0 e^{-\gamma p}$, then

$$\frac{1}{\alpha(p)}\nabla p = \frac{1}{\alpha_0}e^{-\gamma p}\nabla p = -\frac{1}{\alpha_0\gamma}\nabla e^{-\gamma p},$$

and the above equation can be rewritten as

$$-\Delta e^{-\gamma p} = \gamma\nabla\cdot\left(e^{-\gamma p}\mathbf{f}\right). \tag{3.48}$$

Let us introduce the new variable

$$q = e^{-\gamma p} - 1. \tag{3.49}$$

Since $p = 0$ on Γ_w,

$$q = e^{-\gamma p|_{\Gamma_w}} - 1 = 0 \text{ on } \Gamma_w.$$

From (3.48) and (3.49), this new variable satisfies a.e. in Ω

$$-\Delta q - \gamma\nabla\cdot(q\mathbf{f}) = \gamma\nabla\cdot\mathbf{f},$$

$$\alpha(p) = \frac{\alpha_0}{q+1}. \tag{3.50}$$

Assume that the right-hand side $\mathbf{f}$ is smooth enough so that it has a normal trace on Γ. Then it is legitimate to multiply the first equation of (1.1) by $\mathbf{n}$ on Γ and obtain

$$\alpha(p)g + \partial_n p = \mathbf{f}\cdot\mathbf{n}.$$

Denote $\tilde{F} := \mathbf{f}\cdot\mathbf{n}$. By (3.49),

$$\partial_n q + \gamma\tilde{F}q = \alpha_0\gamma g - \gamma\tilde{F}.$$

Thus, for the variable q, we have obtained the following boundary value problem

$$\begin{cases} -\Delta q - \gamma\nabla\cdot(q\mathbf{f}) = \gamma\nabla\cdot\mathbf{f}, & \text{in } \Omega, \\ q = 0, & \text{on } \Gamma_w, \\ \partial_n q + \gamma\tilde{F}q = \alpha_0\gamma g - \gamma\tilde{F}, & \text{on } \Gamma. \end{cases} \tag{3.51}$$

This motivates the following split formulation for problem (1.1):

1. Find q that solves (3.51),

2. In view of (3.50), define

$$\tilde{\alpha}(\mathbf{x}) = \frac{\alpha_0}{q(\mathbf{x})+1}, \quad \mathbf{x}\in\Omega. \tag{3.52}$$

3. Find $(\mathbf{U}, P)$ that solve

$$\begin{cases} \tilde{\alpha}\mathbf{U} + \nabla P = \mathbf{f}, & \text{in } \Omega, \\ \nabla\cdot\mathbf{U} = 0, & \text{in } \Omega, \\ P = p_w, & \text{on } \Gamma_w, \\ \mathbf{U}\cdot\mathbf{n} = g & \text{on } \Gamma. \end{cases} \tag{3.53}$$

Summing up, if $(\mathbf{u}, p)$ is a solution of problem (1.1) and p belongs to $L^\infty(\Omega)$, then $(q, \mathbf{U}, p)$ solves (3.51)–(3.53). The converse is partially established in the next subsection.

Remark 15. This formulation requires only the solution of two linear problems.

Let us first examine the well-posedness of the boundary value problem (3.51). For this, we write it in a variational form. Multiply the first equation of (3.51) by a sufficiently smooth function r that vanishes on Γ_w, apply Green's formula and use the last equation of (3.51). We obtain

$$\int_\Omega \nabla q \cdot \nabla r + \gamma \int_\Omega q\mathbf{f}\cdot\nabla r = \alpha_0\gamma \int_\Gamma gr - \gamma \int_\Omega \mathbf{f}\cdot\nabla r.$$

In the case $d = 3$, the minimal smoothness requirements for these integrals to be meaningful are $q, r \in H^1(\Omega)$, $\mathbf{f} \in \mathbf{L}^3(\Omega)$, and $g \in H_{00}^{1/2}(\Gamma)'$. Hence, the weak formulation of problem (3.51) that we will consider is the following:

Given $\mathbf{f} \in \mathbf{L}^3(\Omega)$ *and* $g \in H_{00}^{1/2}(\Gamma)'$, *find* $q \in H_w^1(\Omega)$ *such that*

$$\int_\Omega \nabla q \cdot r + \gamma \int_\Omega q\mathbf{f}\cdot\nabla r = \alpha_0\gamma \langle g, r\rangle_\Gamma - \gamma \int_\Omega \mathbf{f}\cdot\nabla r, \quad \forall r \in H_w^1(\Omega). \tag{3.54}$$

A sufficient condition for this problem to be well posed is the following.

Proposition 6. *Assume there exists a constant* $\chi < 1$ *such that*

$$\gamma c(\Omega)\, \|\mathbf{f}\|_{\mathbf{L}^3} \le \chi < 1. \tag{3.55}$$

Then, problem (3.54) *has a unique solution* $q \in H_w^1(\Omega)$.

Proof. Let $q = r$ in (3.54); Hölder's inequality and (3.55) give

$$\left|\gamma \int_\Omega q\mathbf{f}\cdot\nabla q\right| \le \gamma \|q\|_{L^6}\|\mathbf{f}\|_{\mathbf{L}^3}\|\nabla q\|_{\mathbf{L}^2} \le \gamma c(\Omega)\|\mathbf{f}\|_{\mathbf{L}^3}|q|_{H^1}^2 \le \chi |q|_{H^1}^2.$$

Then Lax–Milgram's Lemma implies that problem (3.54) is well-posed. □

Remark 16. Condition (3.55) is only sufficient for problem (3.51) to be well-posed. We do not want to provide a thorough analysis of this problem, but only to show that there are cases when the algorithm that we are developing is meaningful.

Next, we turn to problem (3.53). This problem is well-posed if $\tilde{\alpha}$ defined by (3.52) belongs to $L^\infty(\Omega)$ and is bounded away from zero. For this, it suffices that there exists a constant $q_0 > 0$ such that

$$q + 1 \geq q_0 > 0, \text{a.e. in } \Omega, \tag{3.56}$$

and

$$q \in L^\infty(\Omega). \tag{3.57}$$

Condition (3.57) can be regarded as a restriction on the smoothness of the data and the domain. Sufficient conditions for assumption (3.56) to hold elude us at the moment, but we have the following partial result, in the simpler case when $\Gamma_w = \partial\Omega$.

Proposition 7. *Assume that* $\Gamma_w = \partial\Omega$ *and condition* (3.55) *holds. Then* q *satisfies*

$$q + 1 \geq 0, \textit{a.e. in } \Omega.$$

Proof. Let us define the set

$$\Omega^- = \{\mathbf{x} \in \Omega : \ q(\mathbf{x}) + 1 \leq 0\},$$

and the function

$$r_0(\mathbf{x}) = \begin{cases} 0, & \mathbf{x} \notin \Omega^-, \\ -\left(q(\mathbf{x}) + 1\right), & \mathbf{x} \in \Omega^-. \end{cases}$$

Clearly, $r_0 \in H^1(\Omega)$ and by definition $r_0 \geq 0$ almost everywhere in Ω. Moreover, since $q + 1|_{\partial\Omega} = 1 > 0$ then $r_0 \in H_0^1(\Omega)$. By setting $r = r_0$ in (3.54) and changing signs we

obtain that

$$\int_{\Omega^-} |\nabla r_0|^2 + \gamma \int_{\Omega^-} r_0 \mathbf{f} \cdot \nabla r_0 = 0. \tag{3.58}$$

Owing to condition (3.55), equality (3.58) implies that

$$(1 - \chi) \int_{\Omega} |\nabla r_0|^2 \leq 0.$$

In other words $\nabla r_0 = 0$, a.e. in Ω. Since $r_0 \in H_0^1(\Omega)$, we have $r_0 = 0$, a.e. in Ω thus implying the result. □

Under restrictions (3.56), (3.57) and (3.55), we are able to show that the solution $(\mathbf{U}, P)$ to (3.53) solves (1.1).

Proposition 8. *In addition to* (3.55)*, assume that the solution* q *to problem* (3.51) *is in* $L^\infty(\Omega)$ *and satisfies* (3.56)*. Then problem* (3.53) *has a unique solution* $(\mathbf{U}, P)$ *and this solution solves* (1.1)*.*

Proof. By (3.56), there is a unique $\tilde{P}$ such that a.e. in Ω,

$$e^{-\gamma \tilde{P}} = q + 1.$$

The assumption that $q \in L^\infty(\Omega)$ together with (3.56) imply that $\tilde{P} \in H^1(\Omega)$. Moreover, since $q = 0$ on Γ_w, we obtain $\tilde{P} \in H_w^1(\Omega)$.

Define $\tilde{\mathbf{U}} \in \mathbf{L}^2(\Omega)$ by

$$\alpha_0 \gamma \tilde{\mathbf{U}} := \nabla q + \gamma (q + 1) \mathbf{f};$$

by (3.51), this implies that

$$\nabla \cdot \tilde{\mathbf{U}} = 0.$$

Moreover, by the definition of $\tilde{P}$,

$$\alpha_0 \gamma \tilde{\mathbf{U}} = \nabla (e^{-\gamma \tilde{P}} - 1) + \gamma e^{-\gamma \tilde{P}} \mathbf{f} = -\gamma e^{-\gamma \tilde{P}} \nabla \tilde{P} + \gamma e^{-\gamma \tilde{P}} \mathbf{f};$$

hence

$$\alpha(\tilde{P})\tilde{\mathbf{U}} + \nabla\tilde{P} = \mathbf{f}.$$

The boundary condition on $\tilde{\mathbf{U}}$ can be obtained in a similar way. This implies not only that the pair $(\tilde{\mathbf{U}}, \tilde{P})$ solves (1.1), but also that

$$\frac{\alpha_0}{q+1}\tilde{\mathbf{U}} + \nabla\tilde{P} = \mathbf{f}.$$

Since the solution to (3.53) is unique $(\tilde{\mathbf{U}}, \tilde{P}) = (\mathbf{U}, P)$. □

Remark 17. In the case of Dirichlet boundary conditions on the whole boundary: $\Gamma_w = \partial\Omega$, if we slightly restrict the angles of the domain and assume that $\mathbf{f}$ is smoother, for instance $\mathbf{f} \in \mathbf{L}^6(\Omega)$ and $\nabla\cdot\mathbf{f} \in L^2(\Omega)$, then a bootstrap argument, and regularity results for the Laplace equation, show that $q \in W_r^1(\Omega)$ for some $r > 3$ and hence q is continuous. Therefore (3.57) is satisfied.

Let us now discretize (3.51)–(3.53). In order to approximate the linear Darcy system (3.53) we use the spaces $\mathbf{X}_h$ and M_h introduced in Section B and assume that they satisfy (3.17). We also introduce another finite dimensional space $W_h \subset H_w^1(\Omega)$ to discretize (3.51). Then, the discrete algorithm is the following:

1. Find $q_h \in W_h$ such that

$$\int_\Omega \nabla q_h\cdot\nabla s_h + \gamma\int_\Omega q_h\mathbf{f}\cdot\nabla s_h = \alpha_0\gamma\int_\Gamma g s_h - \gamma\int_\Omega \mathbf{f}\cdot\nabla s_h, \quad \forall s_h \in W_h. \tag{3.59}$$

2. Compute the function

$$\tilde{\alpha}_h(\mathbf{x}) = \frac{\alpha_0}{q_h(\mathbf{x})+1}, \quad \mathbf{x}\in\Omega. \tag{3.60}$$

3. Find $(\tilde{\mathbf{u}}_h, \tilde{p}_h) \in \mathbf{X}_h \times M_h$ that solve the discrete linear Darcy system

$$\begin{cases} \int_\Omega \tilde{\alpha}_h \tilde{\mathbf{u}}_h \cdot \mathbf{v}_h + \int_\Omega \mathbf{v}_h \cdot \nabla \tilde{p}_h = \int_\Omega \mathbf{f} \cdot \mathbf{v}_h, & \forall \mathbf{v}_h \in \mathbf{X}_h, \\ \int_\Omega \tilde{\mathbf{u}}_h \cdot \nabla r_h = \langle g, r_h \rangle_\Gamma, & \forall r_h \in M_h. \end{cases} \tag{3.61}$$

Remark 18. Note that finding this approximate solution involves solving only two consecutive linear problems.

Remark 19. Clearly, under assumption (3.55), problem (3.59) has a unique solution. Then, for the discrete version of the splitting method to make sense we need assumptions analogous to (3.56) and (3.57). When W_h has the same structure as in (3.22), (3.57) is always satisfied, although the upper bound may not be uniform with respect to h. Furthermore, if $q_h(\mathbf{x}) + 1 > 0$ for all $\mathbf{x}$ in $\bar{\Omega}$, then since problem (3.61) is set into finite dimension, it also has a unique solution. But of course, (3.56) is not guaranteed, although in the numerical experiments of Section D, we observe indeed that the discrete solution satisfies $q_h + 1 > 0$.

Now, we present an error analysis of the algorithm (3.59)–(3.61), but this analysis is still heuristic because we must assume that the function q_h satisfies uniformly assumptions similar to (3.56) and (3.57). More precisely, we suppose that there are constants $q_{\min}, q_{\max} > 0$ such that for every $h > 0$,

$$0 < q_{\min} \leq q_h(\mathbf{x}) + 1 \leq q_{\max}, \quad \forall \mathbf{x} \in \bar{\Omega}. \tag{3.62}$$

With this, we can proceed in two directions: a straightforward analysis of (3.59)–(3.61), or a comparison with (3.25). In both cases, we suppose that (3.55) holds, so that (3.59) has a unique solution.

Let us proceed first with the second option, namely comparison with (3.25). We do not know whether the nonlinear Darcy problem with exponential porosity has a

solution or not; and if so, which are its properties. For this reason, we shall carry this error analysis under the assumption that problem (1.1) with the function α defined by (1.2) does have a solution. Moreover, we shall assume that the discrete problem defined by (3.25), with α as in (1.2) has a unique solution for all $h > 0$.

Proposition 9. *In addition to* (3.17) *and* (3.55)*, assume that the solution* q_h *to problem* (3.59) *satisfies* (3.62)*. If the pair* $(\tilde{\mathbf{u}}_h, \tilde{p}_h) \in \mathbf{X}_h \times M_h$ *solves* (3.61)*, then there exists a constant* $c > 0$ *independent of* h *such that*

$$\|\mathbf{u}_h - \tilde{\mathbf{u}}_h\|_{\mathbf{L}^2} + |p_h - \tilde{p}_h|_{H^1} \leq c \sup_{\mathbf{x}\in\bar{\Omega}} |\alpha(p_h(\mathbf{x})) - \tilde{\alpha}_h(\mathbf{x})| \|\mathbf{u}_h\|_{\mathbf{L}^2}, \tag{3.63}$$

where $(\mathbf{u}_h, p_h) \in \mathbf{X}_h \times M_h$ *solves* (3.25)*.*

Proof. Let us take the difference of equations (3.25) and (3.61). We obtain

$$\begin{cases} \int_\Omega (\alpha(p_h)\mathbf{u}_h - \tilde{\alpha}_h\tilde{\mathbf{u}}_h)\cdot\mathbf{v}_h + \int_\Omega \mathbf{v}_h\cdot\nabla(p_h - \tilde{p}_h) = 0, & \forall \mathbf{v}_h \in \mathbf{X}_h, \\ \int_\Omega (\mathbf{u}_h - \tilde{\mathbf{u}}_h)\cdot\nabla r_h = 0, & \forall r_h \in M_h. \end{cases}$$

Let $\mathbf{v}_h = \mathbf{u}_h - \tilde{\mathbf{u}}_h$; assumption (3.62) implies

$$\begin{aligned} \frac{\alpha_0}{q_{\max}} \|\mathbf{u}_h - \tilde{\mathbf{u}}_h\|^2_{\mathbf{L}^2} &\leq \int_\Omega \tilde{\alpha}_h |\mathbf{u}_h - \tilde{\mathbf{u}}_h|^2 \\ &= \left| \int_\Omega (\alpha(p_h) - \tilde{\alpha}_h)\, \mathbf{u}_h\cdot(\mathbf{u}_h - \tilde{\mathbf{u}}_h) \right|, \end{aligned}$$

whence

$$\|\mathbf{u}_h - \tilde{\mathbf{u}}_h\|_{\mathbf{L}^2} \leq c \sup_{\mathbf{x}\in\bar{\Omega}} |\alpha(p_h(\mathbf{x})) - \tilde{\alpha}_h(\mathbf{x})| \|\mathbf{u}_h\|_{\mathbf{L}^2}.$$

By the inf–sup condition (3.17)

$$\begin{aligned}\beta|p_h-\tilde{p}_h|_{H^1} &\le \sup_{0\neq\mathbf{v}_h\in\mathbf{X}_h}\frac{\int_\Omega\left(\alpha(p_h)\mathbf{u}_h-\tilde{\alpha}_h\tilde{\mathbf{u}}_h\right)\cdot\mathbf{v}_h}{\|\mathbf{v}_h\|_{\mathbf{L}^2}}\\ &= \sup_{0\neq\mathbf{v}_h\in\mathbf{X}_h}\frac{\int_\Omega\tilde{\alpha}_h\left(\mathbf{u}_h-\tilde{\mathbf{u}}_h\right)\cdot\mathbf{v}_h+\int_\Omega\left(\alpha(p_h)-\tilde{\alpha}_h\right)\mathbf{u}_h\cdot\mathbf{v}_h}{\|\mathbf{v}_h\|_{\mathbf{L}^2}}\\ &\le c\|\mathbf{u}_h-\tilde{\mathbf{u}}_h\|_{\mathbf{L}^2}+\sup_{\mathbf{x}\in\bar{\Omega}}|\alpha(p_h(\mathbf{x}))-\tilde{\alpha}_h(\mathbf{x})|\|\mathbf{u}_h\|_{\mathbf{L}^2}\\ &\le c\sup_{\mathbf{x}\in\bar{\Omega}}|\alpha(p_h(\mathbf{x}))-\tilde{\alpha}_h(\mathbf{x})|\|\mathbf{u}_h\|_{\mathbf{L}^2}.\end{aligned}$$

□

This estimate should be regarded as the basic one. If the exact solution is smooth enough, it can easily be reduced, for instance, to max–norm error estimates for the pressure p and the auxiliary variable q.

Corollary 3. *In addition to* (3.17) *and* (3.55)*, assume that the solution* q *to* (3.54) *belongs to* $L^\infty(\Omega)$ *and satisfies* (3.56)*. Assume, also, that the pair* $(\mathbf{u},p)$ *that solves* (1.1) *is such that* $p\in L^\infty(\Omega)$*. If* q_h *satisfies* (3.62) *then there is a constant* $c>0$ *independent of* h *such that*

$$\|\mathbf{u}_h-\tilde{\mathbf{u}}_h\|_{\mathbf{L}^2}+|p_h-\tilde{p}_h|_{H^1}\le c\left(\|p-p_h\|_{L^\infty}+\|q-q_h\|_{L^\infty}\right)\|\mathbf{u}_h\|_{\mathbf{L}^2}. \tag{3.64}$$

Proof. Using (3.63) it is sufficient to bound the L^∞ norm of the difference $\alpha(p_h)-\tilde{\alpha}_h$. Then

$$\begin{aligned}\|\alpha(p_h)-\tilde{\alpha}_h\|_{L^\infty} &\le \|\alpha(p)-\alpha(p_h)\|_{L^\infty}+\|\alpha(p)-\tilde{\alpha}_h\|_{L^\infty}\\ &\le D\|p-p_h\|_{L^\infty}+\|\alpha(p)-\tilde{\alpha}_h\|_{L^\infty},\end{aligned}$$

where the constant D satisfies

$$D\le\alpha_0\gamma\exp\left(\gamma\max\left\{\|p\|_{L^\infty},\|p_h\|_{L^\infty}\right\}\right).$$

Comparing (3.50) and (3.60), we obtain for a.e. $\mathbf{x}$ in Ω

$$\begin{aligned}|\alpha(p(\mathbf{x})) - \tilde{\alpha}_h(\mathbf{x})| &\leq \alpha_0 \frac{|q_h(\mathbf{x}) - q(\mathbf{x})|}{|(q(\mathbf{x})+1)(q_h(\mathbf{x})+1)|} \\ &\leq \frac{\alpha_0}{|(q(\mathbf{x})+1)(q_h(\mathbf{x})+1)|} \|q_h - q\|_{L^\infty}.\end{aligned}$$

Assumptions (3.56) and (3.62) imply that there is a constant $c > 0$ independent of h such that

$$|(q(\mathbf{x})+1)(q_h(\mathbf{x})+1)| > c \text{ for a.e.} \mathbf{x} \in \Omega,$$

whence (3.64). □

Finally, to be able to provide an order of convergence, we must assume one additional approximation property of the space M_h, and we must assume that the space W_h has adequate approximation properties. More precisely,

1. There is a constant $c > 0$, independent of h, such that for every $r \in W_\infty^\ell(\Omega)$ the interpolation operator $\mathcal{I}_h$ defined in (3.19) satisfies

$$\|r - \mathcal{I}_h r\|_{L^\infty} \leq c h^\ell \|r\|_{W_\infty^\ell}. \tag{3.65}$$

2. There exists an interpolation operator $\rho_h : H^1(\Omega) \to W_h$, such that for all $1 \leq s \leq \infty$, if $r \in W_s^{\ell+1}(\Omega)$

$$\|r - \rho_h r\|_{L^s} + h|r - \rho_h r|_{W_s^1} \leq c h^{\ell+1} \|r\|_{W_s^{\ell+1}}, \tag{3.66}$$

   where the constant $c > 0$ does not depend on r or h.

3. There is a constant $c > 0$ independent of h, such that for every $r_h \in W_h$ the following *inverse inequality* holds

$$\|r_h\|_{L^\infty} \leq c h^{-1/2} |r_h|_{H^1}. \tag{3.67}$$

Remark 20. The space M_h defined in (3.22) has properties (3.65) and (3.66) with the same interpolation operator $\mathcal{I}_h$. Hence, the triple $(\mathbf{X}_h, M_h, M_h)$ with $\mathbf{X}_h$ defined in (3.21) and M_h defined in (3.22) has all the desired properties for all $k \geq 1$.

Under these assumptions, we first bound the error of the auxiliary problem.

Proposition 10. *If* (3.55) *holds, the solution* q_h *of* (3.59) *satisfies*

$$|q - q_h|_{H^1} \leq 2\left(1 + \frac{\gamma c(\Omega)}{1-\chi}\|\mathbf{f}\|_{\mathbf{L}^3}\right) \inf_{r_h \in W_h} |q - r_h|_{H^1}. \tag{3.68}$$

Proof. By taking the difference between (3.59) and (3.54), inserting any function r_h in W_h and testing with $s_h = q_h - r_h$, we obtain

$$|q_h - r_h|_{H^1}\left(1 - \gamma c(\Omega)\|\mathbf{f}\|_{\mathbf{L}^3}\right) \leq |q - r_h|_{H^1}\left(1 + \gamma c(\Omega)\|\mathbf{f}\|_{\mathbf{L}^3}\right).$$

By virtue of (3.55), this implies that

$$|q_h - r_h|_{H^1} \leq \left(1 + 2\frac{\gamma c(\Omega)\|\mathbf{f}\|_{\mathbf{L}^3}}{1 - \gamma c(\Omega)\|\mathbf{f}\|_{\mathbf{L}^3}}\right)|q - r_h|_{H^1}.$$

Then (3.68) follows from (3.55) and the triangle inequality. □

Now we are able to prove a convergence result.

Corollary 4. *In addition to* (3.55)*, assume that the solution* q *to problem* (3.51) *belongs to* $H^{\ell+1}(\Omega) \cap W^{\ell}_{\infty}(\Omega)$ *and satisfies* (3.56)*. Moreover, assume that the solution* $(\mathbf{u}, p)$ *to* (1.1) *is such that* $p \in H^{\ell+1}(\Omega) \cap W^{\ell}_{\infty}(\Omega)$*. Then, if the space* M_h *satisfies* (3.19), (3.35) *and* (3.65)*, and the space* W_h *satisfies* (3.66) *and* (3.67)*, and if* q_h *satisfies* (3.62)*, there exists a constant* $c > 0$ *that does not depend on* h*, such that*

$$\|\mathbf{u}_h - \tilde{\mathbf{u}}_h\|_{\mathbf{L}^2} + |p_h - \tilde{p}_h|_{H^1} \leq ch^{\ell-1/2}\left(|p|_{W^{\ell}_{\infty}} + |p|_{H^{\ell+1}} + |q|_{W^{\ell}_{\infty}} + |q|_{H^{\ell+1}}\right)\|\mathbf{u}_h\|_{\mathbf{L}^2}.$$

Proof. By property (3.65),

$$\|p - p_h\|_{L^\infty} \leq \|p - \mathcal{I}_h p\|_{L^\infty} + \|\mathcal{I}_h p - p_h\|_{L^\infty} \leq ch^\ell |p|_{W^\ell_\infty} + \|\mathcal{I}_h p - p_h\|_{L^\infty}.$$

By the inverse inequality (3.35) and by (3.19)

$$\begin{aligned}\|\mathcal{I}_h p - p_h\|_{L^\infty} &\leq ch^{-1/2} |\mathcal{I}_h p - p_h|_{H^1} \leq ch^{-1/2} \left(|p - \mathcal{I}_h p|_{H^1} + |p - p_h|_{H^1}\right) \\ &\leq ch^{-1/2} \left(h^\ell \|p\|_{H^{\ell+1}} + |p - p_h|_{H^1}\right).\end{aligned}$$

To estimate the term $|p - p_h|_{H^1}$ it is sufficient to recall Corollary 1 in the uniqueness case, or (3.47) for nonsingular solutions (with $s = \ell + 1$). We obtain

$$\|p - p_h\|_{L^\infty} \leq ch^\ell |p|_{W^\ell_\infty} + ch^{\ell - 1/2} |p|_{H^{\ell+1}}.$$

Then we conclude the proof by applying (3.68) and the inverse inequality (3.67). □

Remark 21. The above estimates are suboptimal, but they show heuristically that the splitting algorithm does indeed converge. By using a more refined analysis, for instance the method of weighted norms of Nitsche (see [21], S.C. Brenner and L.R. Scott [15, Chapter 8], or V. Girault, R. Nochetto and L.R. Scott [33], for more details) we may derive (again heuristically) optimal error estimates. The results of Section D give examples where the errors have indeed optimal order.

Remark 22. If q belongs to $H^2(\Omega) \cap W^1_\infty(\Omega)$ and satisfies (3.56), then for all sufficiently small h, q_h also satisfies (3.62).

Now, let us estimate the error of (3.59)–(3.61) without reverting to (3.25). The estimate (3.68) is rigorous because it is derived solely under assumptions on the data. However, the remaining estimates are heuristic because we do not know how to estimate the error on $\tilde{\mathbf{u}}_h$ without assuming that q_h satisfies (3.62) and q satisfies (3.56) and (3.57). Then we have the following result.

Theorem 5. *In addition to* (3.17) *and* (3.55)*, suppose that the solution* q *to* (3.54) *satisfies* (3.56) *and* (3.57)*, the solution* $\mathbf{U}$ *of* (3.53) *belongs to* $\mathbf{L}^3(\Omega)$*, and the solution* q_h *of* (3.59) *satisfies* (3.62)*. Then*

$$\begin{aligned}\|\mathbf{U}-\tilde{\mathbf{u}}_h\|_{\mathbf{L}^2} \leq \left(1+\frac{q_{\max}}{q_{\min}}\right)\left(1+\frac{1}{\beta}\right)\inf_{\mathbf{v}_h\in\mathbf{X}_h}\|\mathbf{U}-\mathbf{v}_h\|_{\mathbf{L}^2} + \frac{q_{\max}}{q_{\min}}\frac{1}{q_0}c(\Omega)\|\mathbf{U}\|_{\mathbf{L}^3}|q-q_h|_{H^1} \\ + \frac{q_{\max}}{\alpha_0}\inf_{r_h\in M_h}|P-r_h|_{H^1},\end{aligned} \tag{3.69}$$

and

$$\begin{aligned}|P-\tilde{p}_h|_{H^1} \leq \left(1+\frac{1}{\beta}\right)\inf_{r_h\in M_h}|P-r_h|_{H^1} + \frac{1}{\beta}\frac{\alpha_0}{q_{\min}}\left(\|\mathbf{U}-\tilde{\mathbf{u}}_h\|_{\mathbf{L}^2}\right. \\ \left. + \frac{c(\Omega)}{q_0}\|\mathbf{U}\|_{\mathbf{L}^3}|q-q_h|_{H^1}\right).\end{aligned} \tag{3.70}$$

Proof. First, the assumptions on q and q_h imply that $\tilde{\alpha}$ and $\tilde{\alpha}_h$ are well-defined and strictly positive. Next, by taking the difference between the first row of (3.61) and (3.53) in weak form, and inserting any element $\mathbf{v}_h$ of $\mathbf{X}_h$ and r_h of M_h, we obtain for any $\mathbf{w}_h$ in $\mathbf{X}_h$,

$$\begin{aligned}\int_\Omega \tilde{\alpha}_h(\tilde{\mathbf{u}}_h-\mathbf{v}_h)\cdot\mathbf{w}_h + \int_\Omega(\tilde{\alpha}_h-\tilde{\alpha})\mathbf{U}\cdot\mathbf{w}_h + \int_\Omega\nabla(\tilde{p}_h-r_h)\cdot\mathbf{w}_h = \\ \int_\Omega\tilde{\alpha}_h(\mathbf{U}-\mathbf{v}_h)\cdot\mathbf{w}_h + \int_\Omega\nabla(P-r_h)\cdot\mathbf{w}_h.\end{aligned}$$

In order to eliminate $\tilde{p}_h$, we proceed as in Theorem 2: owing to (3.17), there exists $\mathbf{v}_h$ in $\mathbf{X}_h$ such that $\mathbf{w}_h := \tilde{\mathbf{u}}_h - \mathbf{v}_h$ belongs to $\mathbf{V}_h$ (see (3.23)), and

$$\|\mathbf{U}-\mathbf{v}_h\|_{\mathbf{L}^2} \leq \left(1+\frac{1}{\beta}\right)\inf_{\mathbf{v}_h\in\mathbf{X}_h}\|\mathbf{U}-\mathbf{v}_h\|_{\mathbf{L}^2}. \tag{3.71}$$

This choice of test function eliminates the last term in the left-hand side of the above

difference. Then by applying (3.62), we derive

$$\|\tilde{\mathbf{u}}_h - \mathbf{v}_h\|_{\mathbf{L}^2} \leq \frac{q_{\max}}{q_{\min}}\|\mathbf{U} - \mathbf{v}_h\|_{\mathbf{L}^2} + \frac{q_{\max}}{\alpha_0}\|\tilde{\alpha}_h - \tilde{\alpha}\|_{L^6}\|\mathbf{U}\|_{\mathbf{L}^3} + \frac{q_{\max}}{\alpha_0}|P - r_h|_{H^1}. \quad (3.72)$$

There remains to estimate $\tilde{\alpha}_h - \tilde{\alpha}$:

$$\|\tilde{\alpha}_h - \tilde{\alpha}\|_{L^6} \leq \frac{\alpha_0}{q_0 q_{\min}} c(\Omega)|q - q_h|_{H^1}. \quad (3.73)$$

Then (3.69) follows by substituting this bound into (3.72) and using (3.71) and the triangle inequality.

To obtain (3.70) notice that, by the discrete inf–sup condition (3.17), for any $r_h \in M_h$

$$\beta|\tilde{p}_h - r_h|_{H^1} \leq \sup_{0\neq\mathbf{y}_h\in\mathbf{X}_h} \frac{b(\mathbf{y}_h, \tilde{p}_h - r_h)}{\|\mathbf{y}_h\|_{\mathbf{L}^2}} \leq |P - r_h|_{H^1} + \sup_{0\neq\mathbf{y}_h\in\mathbf{X}_h} \frac{b(\mathbf{y}_h, P - \tilde{p}_h)}{\|\mathbf{y}_h\|_{\mathbf{L}^2}},$$

which shows that it is sufficient to estimate $b(\mathbf{y}_h, P - \tilde{p}_h)$. By taking the difference of the first equation in (3.53) in weak form and the first equation of (3.61) we obtain

$$\begin{aligned} b(\mathbf{y}_h, P - \tilde{p}_h) &= \int_\Omega (\tilde{\alpha}_h\tilde{\mathbf{u}}_h - \tilde{\alpha}\mathbf{U})\cdot\mathbf{y}_h = \int_\Omega \tilde{\alpha}_h(\tilde{\mathbf{u}}_h - \mathbf{U})\cdot\mathbf{y}_h - \int_\Omega (\tilde{\alpha} - \tilde{\alpha}_h)\mathbf{U}\cdot\mathbf{y}_h \\ &\leq \|\tilde{\alpha}_h\|_{L^\infty}\|\mathbf{U} - \tilde{\mathbf{u}}_h\|_{\mathbf{L}^2}\|\mathbf{y}_h\|_{\mathbf{L}^2} + \|\tilde{\alpha} - \tilde{\alpha}_h\|_{L^6}\|\mathbf{U}\|_{\mathbf{L}^3}\|\mathbf{y}_h\|_{\mathbf{L}^2}, \end{aligned}$$

which, by (3.73) and (3.17) implies

$$|\tilde{p}_h - r_h|_{H^1} \leq \frac{1}{\beta}\left(|P - r_h|_{H^1} + \frac{\alpha_0}{q_{\min}}\left(\|\tilde{\mathbf{u}}_h - \mathbf{U}\|_{\mathbf{L}^2} + \frac{c(\Omega)}{q_0}\|\mathbf{U}\|_{\mathbf{L}^3}|q - q_h|_{H^1}\right)\right). \quad (3.74)$$

The error estimate (3.70) follows from (3.74) and the triangle inequality. □

Remark 23. Proposition 10 and Theorem 5 immediately yield straightforward orders of convergence for $(\tilde{\mathbf{u}}_h, \tilde{p}_h)$. We skip them for the sake of brevity.

D. Numerical Experiments

To illustrate the theory of the previous Sections, we present a series of numerical experiments, in two and three dimensions, which show the performance of the developed methods in a series of testcases.

The numerical experiments in two dimensions were conducted using the package `FreeFem++` (see [53]). In this case, unless otherwise stated, the computational domain is $\Omega = (0,1)^2$, where the top and right sides are Γ_w and the other two sides are Γ.

The numerical experiments in three dimensions were carried out with the help of the `deal.II` library (see [8, 7]). For the experiments in this dimension, the domain is $\Omega = (0,1)^3$, with $\Gamma_w = \{(x,y,z) \in \partial\Omega : \ x = 1\} \cup \{(x,y,z) \in \partial\Omega : \ y = 1\}$ and $\Gamma = \partial\Omega \setminus \bar{\Gamma}_w$.

To test the algorithm (3.30) developed in Section B we have conducted a series of numerical experiments, the results of which we present below. We always initiate the iterative process with $p_h^{(0)} = 0$ and use the stopping criterion

$$\frac{\sqrt{\left\|\mathbf{u}_h^{(n+1)} - \mathbf{u}_h^{(n)}\right\|_{\mathbf{L}^2}^2 + \left|p_h^{(n+1)} - p_h^{(n)}\right|_{H^1}^2}}{\sqrt{\left\|\mathbf{u}_h^{(n+1)}\right\|_{\mathbf{L}^2}^2 + \left|p_h^{(n+1)}\right|_{H^1}^2}} < 10^{-10}.$$

Small Porosity

To test the algorithm in the case when the porosity does not have high variations, we define the porosity as

$$\alpha(\xi) = 1 + \frac{1}{1+\xi^2}, \quad \xi \in \mathbb{R}.$$

Notice that $1 \leq \alpha(\xi) \leq 2$. We define the exact solution as

$$\mathbf{u}(x,y) = (-y^2, z^2, x^2)^\top, \qquad p(x,y) = \sin(2\pi x)\sin(2\pi y)\sin(2\pi z).$$

Table I. 3-D. Iterative Algorithm. Small Porosity. $\mathbb{Q}_1 dc$–velocity, $\mathbb{Q}_1$–pressure.

Level	h	Velocity–$\mathbf{L}^2$	**Rate**	Pressure–H^1	**Rate**	**Iterations**
1	0.5000	1.63E+000	—	3.25E+000	—	5
2	0.2500	9.35E-001	0.80	1.72E+000	0.92	9
3	0.1250	4.97E-001	0.91	8.66E-001	0.99	8
4	0.0625	2.53E-001	0.97	4.35E-001	0.99	8
5	0.0313	1.27E-001	0.99	2.18E-001	1.00	8

These functions determine the right-hand side and boundary data.

The results of the algorithm obtained using a discontinuous-$\mathbb{Q}_1$ approximation of the velocity and a $\mathbb{Q}_1$ approximation of the pressure are reported in Table I. We see that the number of iterations does not depend on the discretization parameter, and the errors on the velocity and pressure have optimal order. We obtained similar results in two dimensions, using spaces $\mathbb{P}_0$-$\mathbb{P}_1$ and $\mathbb{P}_1 dc$-$\mathbb{P}_2$. For the sake of brevity, we do not present them here.

Notice that for the last level the number of cells equals 32,768 and

$$\dim \mathbf{X}_h = 786,432 \qquad \dim M_h = 35,937.$$

Big Porosity

To illustrate the case when the porosity has high variations, but is still bounded we consider

$$\alpha(\xi) = 1 + \frac{10}{1+\xi^2}.$$

Table II. 2-D. Iterative Algorithm. Big Porosity. $\mathbb{P}_1 dc$–velocity, $\mathbb{P}_2$–pressure.

h	Velocity–$\mathbf{L}^2$	**Rate**	Pressure–H^1	**Rate**	**Iterations**
0.250000	2.07E+000	—	9.27E+000	—	14
0.125000	8.57E-001	1.33	2.64E+000	1.43	10
0.062500	2.66E-001	1.27	6.76E-001	1.81	9
0.031250	7.11E-002	1.69	1.69E-001	1.96	9
0.015625	1.81E-002	1.90	4.22E-002	2.00	10

Notice that $1 \leq \alpha(\xi) \leq 11$. We define the exact solution to be

$$\mathbf{u}(x, y) = (-y^2, x^2)^\top, \qquad p(x, y) = 10 \sin(2\pi x) \sin(2\pi y).$$

These functions determine the right-hand side and boundary data.

The results of the algorithm obtained with a discontinuous–$\mathbb{P}_1$ approximation of the velocity and a $\mathbb{P}_2$ approximation of the pressure are reported in Table II. We see that the number of iterations does not depend on the discretization parameter, and the errors on the velocity and pressure have optimal order. Using lower order elements, i.e., a $\mathbb{P}_0$-$\mathbb{P}_1$ approximation, we obtain the same results.

Exponential Porosity

Finally, although the theory developed for algorithm (3.30) does not cover the case of an unbounded (i.e., exponential) porosity, we nevertheless test this case. We set

Table III. 3D Iterative Algorithm. Exponential Porosity. $\mathbb{Q}_1 dc$–velocity, $\mathbb{Q}_1$–pressure.

Level	h	Velocity–$\mathbf{L}^2$	**Rate**	Pressure–H^1	**Rate**	**Iterations**
1	0.5000	3.26E+000	—	3.25E+000	—	8
2	0.2500	1.73E+000	0.91	1.72E+000	0.92	8
3	0.1250	8.93E-001	0.96	8.68E-001	0.98	7
4	0.0625	4.61E-001	0.95	4.39E-001	0.98	7
5	0.0313	2.50E-001	0.88	2.25E-001	0.96	7

the porosity to be defined as in (1.2) with

$$\alpha_0 = 1, \quad \gamma = \frac{1}{4},$$

and the exact solution

$$\mathbf{u}(x,y) = \frac{1}{2}(-y^2, z^2, x^2)^\top, \qquad p(x,y) = 2 + \sin(2\pi x)\sin(2\pi y)\sin(2\pi z).$$

These functions determine the right-hand side and boundary data.

The results of the algorithm obtained using a discontinuous-$\mathbb{Q}_1$ approximation of the velocity and a $\mathbb{Q}_1$ approximation of the pressure are reported in Table III. We see that the number of iterations does not depend on the discretization parameter, and the errors on the velocity and pressure have optimal order. In two dimensions, and on a similar problem, we obtain similar results using $\mathbb{P}_0$-$\mathbb{P}_1$ and $\mathbb{P}_1 dc$-$\mathbb{P}_2$ approximations.

Table IV. 3D Splitting Algorithm. $(\mathbb{Q}_1 dc, \mathbb{Q}_1, \mathbb{Q}_1)$ discretization.

Level	h	Velocity–$\mathbf{L}^2$	**Rate**	Pressure–H^1	**Rate**
1	0.5000	5.25E+000	—	3.25E+000	—
2	0.2500	2.80E+000	0.91	1.72E+000	0.92
3	0.1250	1.45E+000	0.95	8.70E-001	0.98
4	0.0625	7.73E-001	0.91	4.44E-001	0.97
5	0.0313	3.95E-001	0.97	2.35E-001	0.92

Splitting Method

To test the algorithm developed in Section C, let

$$\alpha_0 = 1, \qquad \gamma = \frac{1}{4}.$$

We define the exact solution to be

$$\mathbf{u}(x,y) = \frac{1}{2}(-y^2, z^2, x^2)^\top, \qquad p(x,y) = 2 + \sin(2\pi x)\sin(2\pi y)\sin(2\pi z).$$

Notice that this is the same problem with exponential porosity that we solved using the iterative algorithm. The following triple of finite element spaces was used: $\mathbf{X}_h$–discontinuous-$\mathbb{Q}_1$, M_h–$\mathbb{Q}_1$ and W_h–$\mathbb{Q}_1$. The obtained results can be seen in Table IV. The velocity error in the $\mathbf{L}^2$-norm, and the pressure in the H^1-norm asymptotically have optimal order. Testing the method on a similar two-dimensional problem, we can draw the same conclusions for the triples $(\mathbb{P}_0, \mathbb{P}_1, \mathbb{P}_1)$, $(\mathbb{P}_0, \mathbb{P}_1, \mathbb{P}_2)$, $(\mathbb{P}_1 dc, \mathbb{P}_2, \mathbb{P}_1)$ and $(\mathbb{P}_1 dc, \mathbb{P}_2, \mathbb{P}_2)$.

Table V. 2-D. Computational Time [s]. Exponential Porosity.

	Iterative		**Splitting**			
h	$(\mathbb{P}_0, \mathbb{P}_1)$	$(\mathbb{P}_1 dc, \mathbb{P}_2)$	$(\mathbb{P}_0, \mathbb{P}_1, \mathbb{P}_1)$	$(\mathbb{P}_0, \mathbb{P}_1, \mathbb{P}_2)$	$(\mathbb{P}_1 dc, \mathbb{P}_2, \mathbb{P}_1)$	$(\mathbb{P}_1 dc, \mathbb{P}_2, \mathbb{P}_2)$
5E-1	0.21	0.74	0.02	0.04	0.04	0.06
2.5E-1	0.40	1.13	0.08	0.09	0.10	0.13
1.25E-1	1.20	3.35	0.23	0.27	0.53	0.59
6.25E-2	4.71	23.16	0.95	1.08	5.15	5.25
3.13E-1	23.69	248.62	5.81	7.00	69.87	82.07
1.56E-2	167.36	3341.34	50.64	65.48	1366.66	1702.59
7.81E-3	1711.00	—	713.58	894.86	—	—

Computational Time

In order to estimate the computational complexity of the proposed algorithms, we compare the computational time involved in solving the following two dimensional problem:

$$\alpha(\xi) = e^{\xi/2},$$

$$\mathbf{u} = (-y^3, x^3)^\top, \qquad p(x, y) = 2 + \sin(2\pi x)\sin(2\pi y).$$

We compare the iterative algorithm (3.30) and the splitting method of Section C. The obtained results are shown in Table V.

From the results shown in this Table we can clearly see that the splitting algorithm of Section C outperforms the iterative algorithm (3.30) of Section B. This is expected to be the case, since the splitting algorithm requires solving only two linear

problems as opposed to the iterative algorithm; which although converges independently of the discretization parameter, requires the assembly and solution of a linear problem at each iterative step.

Finally, when comparing the computational times for the splitting algorithm using a fixed velocity-pressure pair but different approximation spaces for the auxiliary problem, we see that the computational times differ very little, their relative difference is never greater than 20%. This suggests that the most time consuming procedure is solving the linear Darcy problem (3.61). This is in agreement with the theory, as this problem has more unknowns and its matrix is indefinite. A better approach for the solution of this problem may reduce the time involved in solving this problem (see the work of J. Schöberl and W. Zulehner [73] and W. Zulehner [87] for instance).

Numerical Investigation of the Convergence Condition for the Iterative Algorithm

In order to further investigate the properties of the iterative algorithm (3.30) and, more precisely, the role of condition (3.31) we solve the following particular problem in the domain

$$\Omega = \left\{(x, y) \in \mathbb{R}^2 : 1 < \sqrt{x^2 + y^2} < 4\right\},$$

with

$$\Gamma_w = \left\{(x, y) \in \mathbb{R}^2 : \sqrt{x^2 + y^2} = 1\right\},$$

and $\Gamma = \partial\Omega \setminus \Gamma_w$. In this domain we solve the nonlinear Darcy equations with exponential porosity. We set the right-hand side that corresponds to the exact solution

$$\mathbf{u}(x, y) = (xr, -yr)^\top, \qquad p(x, y) = r,$$

where $r = \sqrt{x^2 + y^2}$. In the numerical experiments that follow we use a $(\mathbb{P}_0, \mathbb{P}_1, \mathbb{P}_1)$ approximation of the velocity-pressure-auxiliary variable. We set $\alpha_0 = 2$ and vary

the parameter γ. Experimentally we have obtained that if $\gamma < 0.038$ the iterative algorithm converges independently of the initial guess, and it behaves the same way as the cases considered before.

For bigger values of the parameter γ, the splitting algorithm of Section C performs as before. However, the iterative algorithm does not converge anymore. Moreover, if we truncate the porosity function α setting, for instance,

$$\alpha(\xi) = \begin{cases} \alpha_0, & \xi < 0, \\ \alpha_0 e^{\gamma\xi}, & 0 \leq \xi \leq 4.5, \\ \alpha_0 e^{4.5\gamma}, & \xi > 4.5, \end{cases}$$

where the choice of truncation is dictated by $1 \leq p(x, y) \leq 4 \; \forall (x, y) \in \bar{\Omega}$, the method still diverges. For $\gamma = 0.2$, a history of the behavior of the approximate pressure is shown in Figure 1.

From Figure 1 we can see that although the approximate solution diverges, it does remain bounded, and it seems to be oscillating around more than one fixed functions. A detailed analysis of the reasons behind these phenomena is a topic for future research.

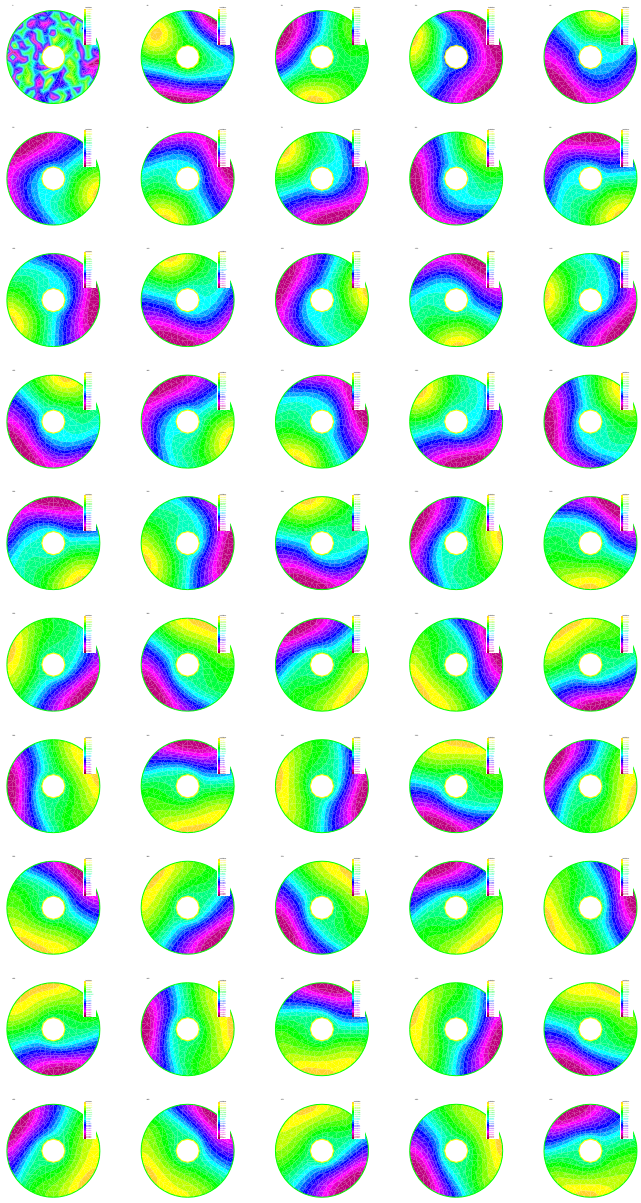

Fig. 1. Approximate pressure for the iterative algorithm. Shown every ten (10) iterations.

CHAPTER IV

THE INCOMPRESSIBLE NAVIER-STOKES EQUATIONS WITH VARIABLE DENSITY *

In this chapter we consider the time-dependent variable density Navier-Stokes system (1.3)–(1.4) on the finite time interval $[0, T]$ and in an open connected and bounded domain $\Omega \subset \mathbb{R}^d$ ($d = 2$ or 3) with boundary $\partial\Omega$, which we assume to be sufficiently smooth. More precisely, we assume that Ω is such that the Stokes operator possesses the usual regularization properties (see [19, 80]). Under these assumptions, our objective is to construct a time and space discretization scheme which has optimal approximation properties and minimizes the computational cost. The space discretization is carried out using Galerkin techniques. The novelty in our approach is the fractional time-stepping technique that we use to discretize in time.

The original results in this chapter were originally presented in [49], [51] and [50]. The organization is as follows. In Section A we review the well known projection schemes for constant density incompressible flows. This proves useful in understanding the difficulties that arise in the case when the density is variable. Moreover, we provide a new proof of a well known result. Namely, the stability of the so-called pressure correction incremental scheme in standard form (see Theorem 9). The novelty in this analysis is that we completely eliminate the projected velocity from the algorithm. Section B presents novel first order schemes for the solution of variable density

---

* Part of the results in this chapter are reprinted with permission from:
A Fractional Step Method Based on a Pressure Poisson Equation for Incompressible Flows with Variable Density by J.-L. Guermond and A. Salgado. C. R. Math. Acad. Sci. Paris, Sér. I 346 (2008), 913–918. Copyright 2008 by Elsevier.
A Splitting Method for Incompressible Flows with Variable Density Based on a Pressure Poisson Equation by J.-L. Guermond and A. Salgado. J. Comput. Phys. 228 (2009), 2834–2846. Copyright 2009 by Elsevier.

incompressible flows. The stability and convergence of these schemes are studied in Section C and Section D, respectively. In Section E a formally second order scheme is introduced and we prove its stability. Finally, Section F presents several numerical experiments that illustrate the performance of the introduced methods.

A. Projection Methods for Constant Density Flows

To understand the ideas and difficulties behind the approximation of variable density flows, let us briefly review the heuristics behind the projection techniques that are used for constant density incompressible flows. For a comprehensive description of these methods we refer the reader to J.-L. Guermond, P.D. Minev and J. Shen [44].

As we stated in Section B of Chapter I, the main difficulty in the approximation of incompressible flows is, in fact, the incompressibility constraint. Let us begin with a technical result, which gives a description of the divergence-free vector fields. For a proof see R. Temam [80, Theorem 1.4].

Theorem 6 (Helmholtz Decomposition)**.** *Let $\Omega \subset \mathbb{R}^d$ be Lipschitz. The following orthogonal decomposition holds*

$$\mathbf{L}^2(\Omega) = \mathbf{H} \oplus \left\{ \mathbf{v} \in \mathbf{L}^2(\Omega) : \ \exists q \in H^1(\Omega) : \mathbf{v} = \nabla q \right\},$$

where

$$\mathbf{H} := \left\{ \mathbf{v} \in \mathbf{L}^2(\Omega) : \nabla \cdot \mathbf{v} = 0, \ \mathbf{v} \cdot \mathbf{n} = 0 \right\}.$$

With this result at hand we can describe the projection methods. To simplify the argumentation, for the time being, let us neglect the nonlinear terms. Moreover, as it is customary in the description of these schemes, we use a semi-discrete setting, i.e., we will not discuss the space discretization. We begin by partitioning the time interval $[0, T]$ into K subintervals, which for the sake of simplicity we take uniform. We then

introduce the time step $\tau = T/N$ and the discrete times $t_k = k\tau$, for $k \in \{0, \ldots, K\}$.

Let us start by reviewing the usual non-incremental Chorin/Temam algorithm for constant density flows [20, 79, 71, 74]. This algorithm for solving the constant density time-dependent Stokes equations consists of computing two sequences of approximate velocities $\{\tilde{\mathbf{u}}^{k+1}\}_{k=0,\ldots,K}$, $\{\mathbf{u}^{k+1}\}_{k=0,\ldots,K}$, and one sequence of approximate pressures $\{p^{k+1}\}_{k=0,\ldots,K}$ as follows: First, set $\mathbf{u}^0 = \mathbf{u}_0$, then for all time steps t_{k+1}, $k \geq 0$, solve

$$\frac{\rho}{\tau}(\tilde{\mathbf{u}}^{k+1} - \mathbf{u}^k) - \mu\Delta\tilde{\mathbf{u}}^k = \mathbf{f}^{k+1}, \qquad \tilde{\mathbf{u}}^{k+1}|_{\partial\Omega} = 0, \tag{4.1}$$

and

$$\frac{1}{\tau}(\mathbf{u}^{k+1} - \tilde{\mathbf{u}}^{k+1}) + \frac{1}{\rho}\nabla p^{k+1} = 0, \qquad \nabla\cdot\mathbf{u}^{k+1} = 0, \qquad \mathbf{u}^{k+1}\cdot\mathbf{n}|_{\partial\Omega} = 0, \tag{4.2}$$

where we have set $\mathbf{f}^{k+1} := \mathbf{f}(t_{k+1})$. One key observation is that, with the help of Theorem 6, the second sub-step can be interpreted as a projection. Indeed, this sub-step can be recast as follows:

$$\mathbf{u}^{k+1} + \frac{\tau}{\rho}\nabla p^{k+1} = \tilde{\mathbf{u}}^{k+1}, \qquad \nabla\cdot\mathbf{u}^{k+1} = 0, \qquad \mathbf{u}^{k+1}\cdot\mathbf{n}|_{\partial\Omega} = 0, \tag{4.3}$$

which is the Helmholtz decomposition of $\tilde{\mathbf{u}}^{k+1}$ into a solenoidal part with zero normal trace plus a gradient. The above decomposition can be equivalently rewritten $\mathbf{u}^{k+1} = P_{\mathbf{H}}\tilde{\mathbf{u}}^{k+1}$, where $P_{\mathbf{H}}$ is the $\mathbf{L}^2$-projection onto $\mathbf{H}$. This fact is the reason this method together with its many avatars is often referred to as a projection algorithm. One very interesting feature of (4.1)–(4.2) is that the pressure can be computed by solving the following Poisson problem:

$$\Delta p^{k+1} = \frac{\rho}{\tau}\nabla\cdot\tilde{\mathbf{u}}^{k+1}, \qquad \partial_n p^{k+1}|_{\partial\Omega} = 0. \tag{4.4}$$

The algorithm (4.1)–(4.2) is simple and can be proved to converge. See e.g. [71, 74, 47] for a proof of the following result.

Theorem 7. *Assume that the solution* (u, p) *to system* (1.3)–(1.4) *is smooth enough. Then, the sequences* $\tilde{\mathbf{u}}_\tau$, $\mathbf{u}_\tau$ *and* p_τ *generated by* (4.1)–(4.2) *satisfy*

$$\|\mathsf{u}_\tau - \tilde{\mathbf{u}}_\tau\|_{\ell^\infty(\mathbf{L}^2)} + \|\mathsf{u}_\tau - \mathbf{u}_\tau\|_{\ell^\infty(\mathbf{L}^2)} \leq c\tau,$$

$$\|\mathsf{u}_\tau - \tilde{\mathbf{u}}_\tau\|_{\ell^2(\mathbf{H}^1)} + \|\mathsf{p}_\tau - p_\tau\|_{\ell^2(L^2)} \leq c\tau^{1/2}.$$

It is important to note at this point that to infer (4.4) from (4.2) we used the fact that the density is constant. When the density is not constant, most of the attempts at splitting the pressure and the velocity that we are aware of so far are based on strategies that are similar to that described above. The main idea always consists of projecting a non-solenoidal provisional velocity onto $\mathbf{H}$; in other words, most of the currently known splitting algorithms consist of solving problems similar to (4.2). When taking the divergence of the left-most equation in (4.2) one is then reduced to solve a problem like the following:

$$-\nabla\cdot\left(\frac{1}{\rho^{k+1}}\nabla\Phi\right) = \Psi, \qquad \partial_n\Phi|_{\partial\Omega} = 0, \tag{4.5}$$

where ρ^{k+1} is an approximation of the non constant function $\rho(t_{n+1})$. It seems that all the algorithms that are more or less based on the Helmholtz decomposition (4.3) always lead to problems like (4.5), which are hard to solve efficiently due to the $1/\rho^{k+1}$ variable coefficient. The key conceptual leap proposed in this dissertation consists of abandoning the projection point of view in favor of a penalty-like argument.

As emphasized in J.-L. Guermond [38] and J.-L. Guermond and L. Quartapelle [46], the projected velocity $\mathbf{u}^{k+1}$ can be eliminated from (4.1)–(4.2). More precisely, the two sub-steps in (4.1)–(4.2) can be equivalently recast as follows:

$$\frac{\rho}{\tau}(\tilde{\mathbf{u}}^{k+1} - \tilde{\mathbf{u}}^k) - \mu\Delta\tilde{\mathbf{u}}^{k+1} + \nabla p^k = \mathbf{f}^{k+1}, \qquad \tilde{\mathbf{u}}^{k+1}|_{\partial\Omega} = 0, \tag{4.6}$$

and

$$\Delta p^{k+1} = \frac{\rho}{\tau}\nabla\cdot\tilde{\mathbf{u}}^{k+1}, \qquad \partial_n p^{k+1}|_{\partial\Omega} = 0. \tag{4.7}$$

Once $\mathbf{u}^{k+1}$ is eliminated, it is clear that the Chorin/Temam algorithm is a discrete version of the following perturbation of the Navier-Stokes equations:

$$\begin{cases} \rho(\mathbf{u}_t + \mathbf{u}\cdot\nabla\mathbf{u}) + \nabla p - \mu\Delta\mathbf{u} = \mathbf{f}, & \mathbf{u}|_{\partial\Omega} = 0, \\ \nabla\cdot\mathbf{u} - \frac{\epsilon}{\rho}\Delta p = 0, & \partial_n p|_{\partial\Omega} = 0, \end{cases} \tag{4.8}$$

where $\epsilon := \tau$. Actually, this perturbation is nothing more than a penalty on the divergence of the velocity as recognized by R. Rannacher in [71], since the momentum equation can also be recast into

$$\rho(\mathbf{u}_t + \mathbf{u}\cdot\nabla\mathbf{u}) + \rho\epsilon^{-1}\nabla\Delta^{-1}\nabla\cdot\mathbf{u} - \mu\Delta\mathbf{u} = \mathbf{f}, \tag{4.9}$$

where Δ^{-1} is the inverse of the Laplace operator equipped with homogeneous Neumann boundary conditions. That is, given $\Psi \in L^2(\Omega)$, we denote by $\Phi = \Delta^{-1}\Psi \in H^1(\Omega)$ the function that has zero mean value and solves

$$\langle\nabla\Phi, \nabla r\rangle = \langle\Psi, r\rangle, \quad \forall r \in H^1(\Omega). \tag{4.10}$$

We shall show that adopting the penalty point of view stated in (4.8) yields efficient splitting algorithms whether the density is constant or not. This point of view is somewhat orthogonal to the current mainstream in the literature which mainly focuses on the projection point of view.

Remark 24. Note that (4.9) is significantly different from standard penalty techniques using $-\epsilon^{-1}\nabla\nabla\cdot\mathbf{u}$ as penalty term, which are generally ill-conditioned. These techniques penalize the divergence in L^2 whereas the term $\epsilon^{-1}\nabla\Delta^{-1}\nabla\cdot\mathbf{u}$ penalize it in a weak norm somewhat related to that of $H^{-1} := (H^1_0)'$.

As we have seen from Theorem 7, the non-incremental pressure correction method is low-order accurate. More precisely, the error is $\mathcal{O}(\tau)$ for the velocity in the $\mathbf{L}^2$-norm and $\mathcal{O}(\tau^{\frac{1}{2}})$ for the velocity in the $\mathbf{H}^1$-norm and the pressure in the L^2-norm. However, Chorin/Temam's constant density algorithm can be improved by making the pressure explicit in the viscous step and by correcting it in the projection step. This technique is usually referred to as the incremental pressure-correction algorithm. This algorithm consists of computing two sequences of approximate velocities $\{\tilde{\mathbf{u}}^{k+1}\}_{k=0,\ldots,K}$, $\{\mathbf{u}^{k+1}\}_{k=0,\ldots,K}$, and one sequence of approximate pressures $\{p^{k+1}\}_{k=0,\ldots,K}$ as follows: First, set $\mathbf{u}^0 = \mathbf{u}_0$, $p^0 = p(0)$, compute an approximation of $\mathbf{u}^1 := \mathbf{u}(\tau)$, then for all time steps t_{k+1}, $k > 1$, solve

$$\frac{\rho}{2\tau}(3\tilde{\mathbf{u}}^{k+1} - 4\mathbf{u}^k + \mathbf{u}^{k-1}) - \mu\Delta\tilde{\mathbf{u}}^{k+1} + \nabla p^k = \mathbf{f}^{k+1}, \qquad \tilde{\mathbf{u}}^{k+1}|_{\partial\Omega} = 0, \tag{4.11}$$

and

$$\frac{3}{2\tau}(\mathbf{u}^{k+1} - \tilde{\mathbf{u}}^{k+1}) + \frac{1}{\rho}\nabla\phi^{k+1} = 0, \qquad \nabla\cdot\mathbf{u}^{k+1} = 0, \qquad \mathbf{u}^{k+1}\cdot\mathbf{n}|_{\partial\Omega} = 0, \tag{4.12}$$

$$p^{k+1} = p^k + \phi^{k+1}. \tag{4.13}$$

Again, the so-called projected velocity (i.e., the solenoidal one) can be algebraically eliminated, thus we obtain the equivalent system

$$\frac{\rho}{2\tau}(3\tilde{\mathbf{u}}^{k+1} - 4\tilde{\mathbf{u}}^k + \tilde{\mathbf{u}}^{k-1}) - \mu\Delta\tilde{\mathbf{u}}^{k+1} + \nabla p^\sharp = \mathbf{f}^{k+1} \quad \tilde{\mathbf{u}}^{k+1}|_{\partial\Omega} = 0, \tag{4.14}$$

$$\Delta\phi^{k+1} = \frac{3\rho}{2\tau}\nabla\cdot\tilde{\mathbf{u}}^{k+1}, \qquad \partial_n\phi|_{\partial\Omega} = 0, \tag{4.15}$$

and (4.13). Here

$$p^\sharp = p^k + \frac{4}{3}\phi^k - \frac{1}{3}\phi^{k-1}. \tag{4.16}$$

The works of J. Shen [74] and J. L. Guermond and L. Quartapelle [46] present an analysis of this scheme. These results can be summarized in the following Theorem.

Theorem 8. *Assume that the solution* (u,p) *to system* (1.3)–(1.4) *is smooth enough. Then, with proper initialization, the sequences* $\tilde{\mathbf{u}}_\tau$, $\mathbf{u}_\tau$ *and* p_τ *generated by* (4.11)–(4.13) *satisfy*

$$\|\mathsf{u}_\tau-\tilde{\mathbf{u}}_\tau\|_{\ell^\infty(\mathbf{L}^2)}+\|\mathsf{u}_\tau-\mathbf{u}_\tau\|_{\ell^\infty(\mathbf{L}^2)}\leq c\tau^2,$$

$$\|\mathsf{u}_\tau-\tilde{\mathbf{u}}_\tau\|_{\ell^2(\mathbf{H}^1)}+\|\mathsf{p}_\tau-p_\tau\|_{\ell^2(L^2)}\leq c\tau.$$

Let us prove stability estimates for this algorithm. As we have seen, this result *per se* is not new but the technique that we use to prove these estimates gives insight on the way to proceed when the density is variable. The main novelty is that the projected velocity is totally eliminated from the analysis. To the best of our knowledge this proof technique has never been used before. This trick enables us to easily extend the proof to the variable density case (see Section E). To avoid irrelevant technicalities assume that $\mathbf{f}\equiv 0$. We now prove that algorithm (4.14)–(4.15) and (4.13) is stable.

Theorem 9. *Let* $\rho\equiv 1$. *The solution* $\{(\tilde{\mathbf{u}}^k,p^k)\}_{k\geq 0}$ *to* (4.14)–(4.15) *and* (4.13) *satisfies the following estimate:*

$$\begin{aligned}\|\tilde{\mathbf{u}}^k\|_{\mathbf{L}^2}^2+\tau^2\|\nabla p^k\|_{\mathbf{L}^2}^2+\tau^2\|\nabla\delta p^{k-1}\|_{\mathbf{L}^2}^2+\sum_{l=2}^{k}\left[\tau\|\tilde{\mathbf{u}}^l\|_{\mathbf{H}^1}^2+\tau^2\|\nabla\delta p^{l-1}\|_{\mathbf{L}^2}^2\right]\\ \leq c\left(\|\tilde{\mathbf{u}}^0\|_{\mathbf{L}^2}^2+\|\tilde{\mathbf{u}}^1\|_{\mathbf{L}^2}^2+\tau^2\|\nabla p^0\|_{\mathbf{L}^2}^2+\tau^2\|\nabla p^1\|_{\mathbf{L}^2}^2\right),\quad\forall k\geq 2.\end{aligned}$$

Proof. We proceed in two steps:

(i) *Initialization:* We consider the steps $k=1,2$ separately as they involve the initial quantities. Let us begin by noticing that the definition of $p^\sharp$ involves only terms from the previous time steps. For $k=1$ or 2 multiply (4.14) by $4\tau\tilde{\mathbf{u}}^{k+1}$. Using the identity

$$2a^{k+1}\left(3a^{k+1}-4a^k+a^{k-1}\right)=\left|a^{k+1}\right|^2+\left|2a^{k+1}-a^k\right|^2+\left|\delta^2a^{k+1}\right|^2-\left|a^k\right|^2-\left|2a^k-a^{k-1}\right|^2,$$

and the Cauchy-Schwarz inequality we obtain

$$\frac{1}{2}\|\tilde{\mathbf{u}}^{k+1}\|_{\mathbf{L}^2}^2+\|2\tilde{\mathbf{u}}^{k+1}-\tilde{\mathbf{u}}^{k}\|_{\mathbf{L}^2}^2+4\tau\|\tilde{\mathbf{u}}^{k+1}\|_{\mathbf{H}^1}^2\leq\|\tilde{\mathbf{u}}^{k}\|_{\mathbf{L}^2}^2+\|2\tilde{\mathbf{u}}^{k}-\tilde{\mathbf{u}}^{k-1}\|_{\mathbf{L}^2}^2+8\tau^2\|\nabla p^{\sharp}\|_{\mathbf{L}^2}^2,$$

which implies

$$\|\tilde{\mathbf{u}}^{k+1}\|_{\mathbf{L}^2}^2+\tau\|\tilde{\mathbf{u}}^{k+1}\|_{\mathbf{H}^1}^2\leq c\left(\|\tilde{\mathbf{u}}^{0}\|_{\mathbf{L}^2}^2+\|\tilde{\mathbf{u}}^{1}\|_{\mathbf{L}^2}^2+\tau^2\|\nabla p^{0}\|_{\mathbf{L}^2}^2+\tau^2\|\nabla p^{1}\|_{\mathbf{L}^2}^2\right),$$

for $k=1$ or 2. The estimate on the pressure is obtained by eliminating ϕ^{k+1} from (4.15) using (4.13), multiplying by δp^{k+1} and integrating by parts. Again, the Cauchy-Schwarz inequality implies

$$\frac{4\tau^2}{9}\|\nabla\delta p^{k+1}\|_{\mathbf{L}^2}^2\leq\|\tilde{\mathbf{u}}^{k+1}\|_{\mathbf{L}^2}^2.$$

The triangle inequality and the estimates obtained above imply the claimed estimate for the first two steps $k=1,2$.

(ii) *General Step:* For $k\geq 3$ notice that, by (4.13)

$$p^{\sharp}=\frac{7p^{k}-5p^{k-1}+p^{k-2}}{3}=\frac{3p^{k+1}-3\delta^2p^{k+1}+\delta^2p^{k}}{3}.$$

Multiply (4.14) by $4\tau\tilde{\mathbf{u}}^{k+1}$ and integrate. Using the identity

$$\begin{aligned}2a^{k+1}\left(3a^{k+1}-4a^{k}+a^{k-1}\right)=3\left|a^{k+1}\right|^2-4\left|a^{k}\right|^2+\left|a^{k-1}\right|^2\\+2\left|\delta a^{k+1}\right|^2-2\left|\delta a^{k}\right|^2+\left|\delta^2a^{k+1}\right|^2,\end{aligned}$$

we obtain

$$\begin{aligned}&3\|\tilde{\mathbf{u}}^{k+1}\|_{\mathbf{L}^2}^2-4\|\tilde{\mathbf{u}}^{k}\|_{\mathbf{L}^2}^2+\|\tilde{\mathbf{u}}^{k-1}\|_{\mathbf{L}^2}^2\\&\quad+2\|\delta\tilde{\mathbf{u}}^{k+1}\|_{\mathbf{L}^2}^2-2\|\delta\tilde{\mathbf{u}}^{k}\|_{\mathbf{L}^2}^2+\|\delta^2\tilde{\mathbf{u}}^{k+1}\|_{\mathbf{L}^2}^2+4\tau\|\tilde{\mathbf{u}}^{k+1}\|_{\mathbf{H}^1}^2\\&\quad+4\tau\left\langle\nabla p^{k+1},\tilde{\mathbf{u}}^{k+1}\right\rangle-4\tau\left\langle\nabla\delta^2p^{k+1},\tilde{\mathbf{u}}^{k+1}\right\rangle+\frac{4\tau}{3}\left\langle\nabla\delta^2p^{k},\tilde{\mathbf{u}}^{k+1}\right\rangle=0.\end{aligned}$$

From the projection equation (4.15) and the pressure update equation (4.13) we deduce that

$$\left\langle \nabla r, \tilde{\mathbf{u}}^{k+1} \right\rangle = \frac{2\tau}{3} \left\langle \nabla r, \nabla \delta p^{k+1} \right\rangle, \quad \forall r \in H^1(\Omega).$$

Using this property together with the identity $2a(a-b) = a^2 - b^2 + (a-b)^2$ we infer

$$\begin{aligned} 3\|\tilde{\mathbf{u}}^{k+1}\|_{\mathbf{L}^2}^2 - 4\|\tilde{\mathbf{u}}^{k}\|_{\mathbf{L}^2}^2 + \|\tilde{\mathbf{u}}^{k-1}\|_{\mathbf{L}^2}^2 + 2\|\delta\tilde{\mathbf{u}}^{k+1}\|_{\mathbf{L}^2}^2 - 2\|\delta\tilde{\mathbf{u}}^{k}\|_{\mathbf{L}^2}^2 \\ + \|\delta^2\tilde{\mathbf{u}}^{k+1}\|_{\mathbf{L}^2}^2 + 4\tau\|\tilde{\mathbf{u}}^{k+1}\|_{\mathbf{H}^1}^2 + \frac{4\tau^2}{3}\left[\|\nabla p^{k+1}\|_{\mathbf{L}^2}^2 - \|\nabla p^{k}\|_{\mathbf{L}^2}^2 \right. \\ \left. + \|\nabla \delta p^{k}\|_{\mathbf{L}^2}^2 - \|\nabla \delta^2 p^{k+1}\|_{\mathbf{L}^2}^2\right] + \frac{8\tau^2}{9}\left\langle \nabla\delta^2 p^k, \nabla\delta p^{k+1}\right\rangle = 0. \end{aligned}$$

Now we use the following identity:

$$\|\delta\tilde{\mathbf{u}}\|_{\mathbf{L}^2}^2 = \|\delta\tilde{\mathbf{u}} - \frac{2\tau}{3}\nabla\delta^2 p\|_{\mathbf{L}^2}^2 + \frac{4\tau^2}{9}\|\nabla\delta^2 p\|_{\mathbf{L}^2}^2$$

which we apply at time steps t_{k+1} and t_k (note that it is critical to have $k \geq 3$ here) and we obtain

$$\begin{aligned} 3\|\tilde{\mathbf{u}}^{k+1}\|_{\mathbf{L}^2}^2 - 4\|\tilde{\mathbf{u}}^{k}\|_{\mathbf{L}^2}^2 + \|\tilde{\mathbf{u}}^{k-1}\|_{\mathbf{L}^2}^2 + \|\delta^2\tilde{\mathbf{u}}^{k+1}\|_{\mathbf{L}^2}^2 + 4\tau\|\tilde{\mathbf{u}}^{k+1}\|_{\mathbf{H}^1}^2 \\ + 2\|\delta\tilde{\mathbf{u}}^{k+1} - \frac{2\tau}{3}\nabla\delta^2 p^{k+1}\|_{\mathbf{L}^2}^2 - 2\|\delta\tilde{\mathbf{u}}^{k} - \frac{2\tau}{3}\nabla\delta^2 p^{k}\|_{\mathbf{L}^2}^2 \\ + \frac{4\tau^2}{3}\left[\|\nabla p^{k+1}\|_{\mathbf{L}^2}^2 - \|\nabla p^{k}\|_{\mathbf{L}^2}^2 + \|\nabla\delta p^{k}\|_{\mathbf{L}^2}^2\right] \\ - \frac{4\tau^2}{9}\|\nabla\delta^2 p^{k+1}\|_{\mathbf{L}^2}^2 - \frac{8\tau^2}{9}\|\nabla\delta^2 p^{k}\|_{\mathbf{L}^2}^2 + \frac{8\tau^2}{9}\left\langle \nabla\delta^2 p^k, \nabla\delta p^{k+1}\right\rangle = 0. \end{aligned}$$

We observe from this inequality that we need to control the last three terms. We rewrite these as follows:

$$\begin{aligned} -\frac{4\tau^2}{9}\|\nabla\delta^2 p^{k+1}\|_{\mathbf{L}^2}^2 - \frac{8\tau^2}{9}\|\nabla\delta^2 p^{k}\|_{\mathbf{L}^2}^2 + \frac{8\tau^2}{9}\left\langle \nabla\delta^2 p^k, \nabla\delta p^{k+1}\right\rangle = \\ -\frac{4\tau^2}{9}\|\nabla\delta^3 p^{k+1}\|_{\mathbf{L}^2}^2 - \frac{4\tau^2}{9}\left\langle \nabla\delta^2 p^k, \nabla\left(\delta^2 p^k + 2\delta^2 p^{k+1} - 2\delta p^{k+1}\right)\right\rangle. \end{aligned}$$

Use (4.13) to eliminate ϕ^{k+1} from (4.15). Applying δ^2 to the result, multiplying it by

$\delta^3 p^{k+1}$, integrating and using the Cauchy-Schwarz inequality we obtain, for $k \geq 3$

$$\frac{4\tau^2}{9}\|\nabla\delta^3 p^{k+1}\|^2_{\mathbf{L}^2} \leq \|\delta^2\tilde{\mathbf{u}}^{k+1}\|^2_{\mathbf{L}^2}.$$

Observing that $\delta^2 p^k + 2\delta^2 p^{k+1} - 2\delta p^{k+1} = -\delta p^k - \delta p^{k-1}$ and using the inequality above, we obtain the following bound:

$$-\frac{4\tau^2}{9}\|\delta^2 p^{k+1}\|^2 - \frac{8\tau^2}{9}\|\delta^2 p^k\|^2 + \frac{8\tau^2}{9}\left\langle\nabla\delta^2 p^k, \nabla\delta p^{k+1}\right\rangle \geq \\ -\|\delta^2\tilde{\mathbf{u}}^{k+1}\|^2_{\mathbf{L}^2} + \frac{4\tau^2}{9}\left[\|\delta p^k\|^2 - \|\delta p^{k-1}\|^2\right],$$

from which we finally deduce the following energy estimate:

$$\begin{aligned} &3\|\tilde{\mathbf{u}}^{k+1}\|^2_{\mathbf{L}^2} - 4\|\tilde{\mathbf{u}}^k\|^2_{\mathbf{L}} + \|\tilde{\mathbf{u}}^{k-1}\|^2_{\mathbf{L}^2} + 4\tau\|\tilde{\mathbf{u}}^{k+1}\|^2_{\mathbf{H}^1} \\ &\qquad + 2\|\delta\tilde{\mathbf{u}}^{k+1} - \frac{2\tau}{3}\nabla\delta^2 p^{k+1}\|^2_{\mathbf{L}^2} - 2\|\delta\tilde{\mathbf{u}}^k - \frac{2\tau}{3}\nabla\delta^2 p^k\|^2_{\mathbf{L}^2} \\ &+ \frac{4\tau^2}{3}\left[\|\nabla p^{k+1}\|^2_{\mathbf{L}^2} - \|\nabla p^k\|^2_{\mathbf{L}^2} + \|\nabla\delta p^k\|^2_{\mathbf{L}^2}\right] + \frac{4\tau^2}{9}\left[\|\nabla\delta p^k\|^2_{\mathbf{L}^2} - \|\nabla\delta p^{k-1}\|^2_{\mathbf{L}^2}\right] \leq 0. \end{aligned} \tag{4.17}$$

We are now going to use the stability estimates proved in Appendix A. Let us define the quantities

$$\begin{aligned} a^s &:= \|\tilde{\mathbf{u}}^s\|^2_{\mathbf{L}^2}, \\ b^s &:= 4\tau\|\tilde{\mathbf{u}}^s\|^2_{\mathbf{H}^1} + \frac{4\tau^2}{3}\|\nabla\delta p^{s-1}\|^2_{\mathbf{L}^2}, \\ d^s &:= 2\|\delta\tilde{\mathbf{u}}^s - \frac{2\tau}{3}\nabla\delta^2 p^s\|^2_{\mathbf{L}^2} + \frac{4\tau^2}{3}\|\nabla p^s\|^2_{\mathbf{L}^2} + \frac{4\tau^2}{9}\|\nabla\delta p^{s-1}\|^2_{\mathbf{L}^2}. \end{aligned}$$

Then (4.17) can be rewritten as

$$3a^{k+1} - 4a^k + a^{k-1} \leq -\left(b^{k+1} + d^{k+1} - d^k\right), \quad k \geq 3$$

Setting $g^{k+1} := -\left(b^{k+1} + d^{k+1} - d^k\right)$ this three-term recursion inequality satisfies the

hypotheses of Corollary 6 (see Appendix A) for $k \geq 3$. Hence

$$a^{\nu} \leq c\left(a^1 + a^2\right) - \sum_{l=3}^{\nu} \frac{1}{3^{\nu+1-l}} \sum_{s=3}^{l} \left(b^s + d^s - d^{s-1}\right), \quad \nu \geq 3$$

or

$$a^{\nu} + \sum_{l=3}^{\nu} \frac{1}{3^{\nu+1-l}} d^l + \sum_{l=3}^{\nu} \frac{1}{3^{\nu+1-l}} \sum_{s=3}^{l} b^s \leq c\left(a^1 + a^2 + d^2\right), \quad \nu \geq 3.$$

Dropping some positive terms in the left-hand side we deduce

$$a^{\nu} + \frac{1}{3} d^{\nu} + \frac{1}{3} \sum_{s=2}^{\nu} b^s \leq c\left(a^1 + a^2 + d^2\right).$$

Given the bounds obtained in the initialization step, this inequality implies the claimed result. □

If, in (4.14)–(4.15), the difference quotients are replaced by time derivatives and the remaining τ's are replaced by ϵ, the above algorithm reduces to the following perturbation of the Navier-Stokes equations:

$$\begin{cases} \rho(\mathbf{u}_t + \mathbf{u}\cdot\nabla\mathbf{u}) + \nabla p - \mu\Delta\mathbf{u} = \mathbf{f}, & \mathbf{u}|_{\partial\Omega} = 0, \\ \nabla\cdot\mathbf{u} - \frac{\epsilon}{\rho}\Delta\phi = 0, & \partial_n\phi|_{\partial\Omega} = 0, \\ \epsilon p_t = \phi. \end{cases} \tag{4.18}$$

Formally, (4.18) is a $\mathcal{O}(\epsilon^2)$ perturbation of the constant density incompressible Navier-Stokes equations. The system (4.18) serves as the starting point for the new algorithm for variable density flows developed in Section E.

Remark 25. L.J.P. Timmermans et al. [82] proposed another version of this scheme which, following the terminology of [44] is called rotational. In this version, the pressure update step (4.13) is replaced by

$$p^{k+1} = p^k + \phi^{k+1} - \mu\nabla\cdot\tilde{\mathbf{u}}^{k+1}. \tag{4.19}$$

In this case, we can see that the scheme corresponds to the following perturbation of the Navier-Stokes equations:

$$\begin{cases} \rho(\mathbf{u}_t + \mathbf{u}\cdot\nabla\mathbf{u}) + \nabla p - \mu\Delta\mathbf{u} = \mathbf{f}, & \mathbf{u}|_{\partial\Omega} = 0, \\ \nabla\cdot\mathbf{u} - \frac{\epsilon}{\rho}\Delta\phi = 0, & \partial_n\phi|_{\partial\Omega} = 0, \\ \epsilon p_t = \phi - \frac{\mu}{\rho}\nabla\cdot\mathbf{u}. \end{cases}$$

Moreover, J.L Guermond and J. Shen have shown (cf. [52]) that this divergence correction significantly improves the pressure approximation. To be more precise, the velocity error in the $\ell^2(\mathbf{H}^1)$-norm and the pressure in the $\ell^2(L^2)$-norm are $\mathcal{O}(\tau^{3/2})$. This is the best convergence result known so far.

B. Description of the First Order Schemes

On the basis of the observations of the previous section, we are going to construct a fractional time-stepping technique for incompressible flows with variable density. Let us begin by describing the space discretization. To construct a Galerkin approximation of (1.3)–(1.4) we introduce three sequences of finite-dimensional spaces $\{W_h\}_{h>0}$, $\{\mathbf{X}_h\}_{h>0}$, $\{M_h\}_{h>0}$, for $h > 0$, with $W_h \subset H^1(\Omega)$, $\mathbf{X}_h \subset \mathbf{H}^1_0(\Omega)$ and $M_h \subset H^1(\Omega)$. We use W_h, $\mathbf{X}_h$, and M_h to approximate the density, the velocity, and the pressure, respectively. With these spaces we can now describe the first-order fractional time-stepping schemes.

Initialization Given the initial data $(\rho_0, \mathbf{u}_0)$, we construct the approximate data $(\rho_h^0, \mathbf{u}_h^0, p_h^0) \in W_h \times \mathbf{X}_h \times M_h$. The initial pressure p_h^0 can be computed from the pair $(\rho_0, \mathbf{u}_0)$, see [44] for more details.

Time-Stepping Technique Given $(\rho_h^k, \mathbf{u}_h^k, p_h^k) \in W_h \times \mathbf{X}_h \times M_h$ we now describe how to obtain the next approximations $(\rho_h^{k+1}, \mathbf{u}_h^{k+1}, p_h^{k+1}) \in W_h \times \mathbf{X}_h \times M_h$. The

algorithm proceeds in three steps: density update, velocity update, pressure update.

Density Update The density update is computed using the mass conservation equation, which we recall is hyperbolic. It is well known that Galerkin techniques are not well suited for the solution of hyperbolic problems (see for instance [27]). The list of techniques aiming at addressing this issue is endless; among these methods one can cite Galerkin-Least Squares [57], Discontinuous-Galerkin [57, 85], subgrid viscosity [40], method of characteristics [25], edge stabilization [18], entropy viscosity [45] and many others. We assume that the sequence of approximate densities $\{\rho_h^k\}_{k=0,\dots,K} \subset W_h$ is obtained by one of these stabilization techniques. More precisely, we assume that given the pair $(\rho_h^k, \mathbf{u}_h^k) \in W_h \times \mathbf{X}_h$, the approximation technique that is used to approximate the mass conservation returns ρ_h^{k+1} and that this algorithm satisfies the following stability hypothesis:

$$\chi \leq \min_{\mathbf{x}\in\bar{\Omega}} \rho_h^{k+1}, \qquad \sup_{\mathbf{x}\in\bar{\Omega}} \rho_h^{k+1} \leq \varrho, \qquad \forall k \geq 1. \tag{4.20}$$

Note that this is a natural assumption since, owing to the incompressibility of the velocity field, the density field ρ satisfies the following property:

$$\rho(t) \in \left[\min_{\mathbf{x}\in\bar{\Omega}} \rho_0(x), \sup_{\mathbf{x}\in\bar{\Omega}} \rho_0(x)\right],$$

for all $t \geq 0$, cf. [61]. For instance, first-order monotone schemes satisfy (4.20) with $\chi = \min_{\mathbf{x}\in\bar{\Omega}} \rho_0(x)$ and $\varrho = \sup_{\mathbf{x}\in\bar{\Omega}} \rho_0(x)$.

Velocity Update Having obtained an approximate density, we define

$$\rho_h^\star := \frac{1}{2}\left(\rho_h^{k+1} + \rho_h^k\right), \tag{4.21}$$

$$p_h^\sharp := p_h^k + \gamma\delta p_h^k, \quad \gamma \in \{0, 1\}. \tag{4.22}$$

The parameter γ is user-dependent. We say that the method is non-incremental if $\gamma = 0$ and incremental if $\gamma = 1$. The incremental version of the algorithm is more accurate that the non-incremental one. We consider the non-incremental version of the algorithm for historical reasons. As we have seen above, the original algorithm of Chorin and Temam for constant density incompressible flows is non-incremental. When $\gamma = 1$, we take $\delta p_h^0 = 0$. The next approximation of the velocity field $\mathbf{u}_h^{k+1} \in \mathbf{X}_h$ is computed by solving the following problem:

$$\left\langle \frac{\rho_h^\star \mathbf{u}_h^{k+1} - \rho_h^k \mathbf{u}_h^k}{\tau}, \mathbf{v}_h \right\rangle + \left\langle \rho_h^{k+1} \mathbf{u}_h^k \cdot \nabla \mathbf{u}_h^{k+1}, \mathbf{v}_h \right\rangle + \left\langle \tfrac{1}{2} \nabla \cdot \left(\rho_h^{k+1} \mathbf{u}_h^k \right) \mathbf{u}_h^{k+1}, \mathbf{v}_h \right\rangle$$
$$+ \mu \left\langle \nabla \mathbf{u}_h^{k+1}, \nabla \mathbf{v}_h \right\rangle + \left\langle \nabla p_h^\sharp, \mathbf{v}_h \right\rangle = \left\langle f^{k+1}, \mathbf{v}_h \right\rangle, \ \forall \mathbf{v}_h \in \mathbf{X}_h. \quad (4.23)$$

Penalty We let $\phi_h^\flat \in M_h$ be the solution of:

$$\left\langle \nabla \phi_h^\flat, \nabla r_h \right\rangle = \frac{\chi}{\tau} \left\langle \mathbf{u}_h^{k+1}, \nabla r_h \right\rangle, \ \forall r_h \in M_h, \quad (4.24)$$

Pressure Update Finally, we define the pressure approximation $p_h^{k+1} \in M_h$ by

$$p_h^{k+1} = \phi_h^\flat + \gamma p_h^k, \quad \gamma \in \{0, 1\}. \quad (4.25)$$

Remark 26. The term $\frac{1}{\tau}[\rho_h^\star \mathbf{u}_h^{k+1} - \rho_h^k \mathbf{u}_h^k] + \frac{1}{2} \nabla \cdot (\rho_h^{k+1} \mathbf{u}_h^k) \mathbf{u}_h^{k+1}$ in (4.23) is asymptotically consistent with the equation. Notice that if the involved functions are sufficiently smooth

$$\begin{aligned} \frac{\frac{1}{2}(\rho_h^{k+1} + \rho_h^k)\mathbf{u}_h^{k+1} - \rho_h^k \mathbf{u}_h^k}{\tau} + \frac{1}{2} \nabla \cdot \left(\rho_h^{k+1} \mathbf{u}_h^k \right) \mathbf{u}_h^{k+1} &= \rho_h^k \frac{\mathbf{u}_h^{k+1} - \mathbf{u}_h^k}{\tau} \\ &+ \frac{(\rho_h^{k+1} - \rho_h^k)}{2\tau} \mathbf{u}_h^{k+1} + \frac{1}{2} \nabla \cdot (\rho_h^{k+1} \mathbf{u}_h^k) \mathbf{u}_h^{k+1} \\ &= [\rho_h (\mathbf{u}_h)_t]^{k+1} + \mathcal{O}(\tau), \end{aligned}$$

The purpose of this particular way of discretizing the quantity $\rho \mathsf{u}_t$ will become clear

once we do the stability analysis.

Remark 27. Let us introduce the auxiliary space $\mathbf{Y}_h := \mathbf{X}_h + \nabla M_h$. In view of (4.24), the quantity

$$\bar{\mathbf{u}}_h^k := \mathbf{u}_h^k - \frac{\tau}{\chi}\nabla\phi_h^\flat \in \mathbf{Y}_h,$$

is discretely divergence free (in the sense that $\langle \bar{\mathbf{u}}_h^k, \nabla r_h\rangle = 0$ for all $r_h \in M_h$) and could be used as a solenoidal approximation of the velocity. This particular choice of $\mathbf{Y}_h$ fits into the commutative diagram framework described in [37, 38, 47]. Therefore, it could be possible to develop a much more general theory about fractional time-stepping techniques for variable density incompressible flows that would include our method as a particular instance. More specifically, let us assume that one has at hand a space $\mathbf{Y}_h$ so that $\mathbf{X}_h \subset \mathbf{Y}_h$. Let $B_h : \mathbf{X}_h \to M_h$ be the operator defined by $\langle B_h\mathbf{v}_h, q_h\rangle := \langle \nabla\cdot\mathbf{v}_h, q_h\rangle$ for all $\mathbf{v}_h \in \mathbf{X}_h$ and all q_h in M_h. Assume that one can construct an extension of B_h over $\mathbf{Y}_h$, say $C_h : \mathbf{Y}_h \to M_h$. The operator C_h being an extension of B_h over $\mathbf{Y}_h$ means that $B_h = C_h i_h$, where i_h is the natural injection $i_h : \mathbf{X}_h \longrightarrow \mathbf{Y}_h$. Then, in this setting, our theory will work by replacing (4.24) by

$$C_h C_h^T \phi_h^\flat = \frac{\chi}{\tau} B_h \mathbf{u}_h^{k+1}. \tag{4.26}$$

For the sake of clarity, we shall not pursue this direction. However, the reader can easily verify that the arguments presented here extend to this situation.

C. Stability of the First-Order Schemes

To obtain stability estimates we henceforth assume that $\min_{\mathbf{x}\in\bar{\Omega}} \rho_0(x) > 0$ (i.e., that there is no vacuum), and that the sequence of approximate density fields $\{\rho_h^k\}$ satisfies property (4.20). Moreover, to avoid irrelevant technicalities, we assume that there is no driving force, i.e., $\mathbf{f} \equiv 0$. Under this assumptions the stability of the non-

incremental scheme is given by the following Theorem.

Theorem 10. *Assume that* (4.20) *holds. Then, for any* $\tau > 0$ *the solution* $\mathbf{u}_{h,\tau} \subset \mathbf{X}_h$ *and* $p_{h,\tau} \subset M_h$ *to the scheme of Section B with* $\gamma = 0$ *satisfies the following inequality:*

$$\|\sigma_h^K \mathbf{u}_h^K\|_{\mathbf{L}^2}^2 + 2\mu\tau \sum_{k=1}^{K} \|\nabla \mathbf{u}_h^k\|_{\mathbf{L}^2}^2 + \frac{\tau^2}{\chi} \sum_{k=0}^{K} \|\nabla p_h^k\|_{\mathbf{L}^2}^2 \leq \|\sigma_h^0 \mathbf{u}_h^0\|_{\mathbf{L}^2}^2,$$

where $\sigma_h^k := \sqrt{\rho_h^k}$.

Proof. We begin by setting $\mathbf{v}_h = 2\tau \mathbf{u}_h^{k+1}$ in the momentum equation (4.23). Notice then that

$$2\left\langle \tfrac{1}{2}(\rho_h^{k+1} + \rho_h^k)\mathbf{u}_h^{k+1} - \rho_h^k \mathbf{u}_h^k, \mathbf{u}_h^{k+1} \right\rangle = \|\sigma_h^{k+1}\mathbf{u}_h^{k+1}\|_{\mathbf{L}^2}^2 + \|\sigma_h^k \delta \mathbf{u}_h^{k+1}\|_{\mathbf{L}^2}^2 - \|\sigma_h^k \mathbf{u}_h^l\|_{\mathbf{L}^2}^2.$$

Moreover, given the boundary conditions

$$\left\langle \rho_h^{k+1}\mathbf{u}_h^k \cdot \nabla \mathbf{u}_h^{k+1} + \tfrac{1}{2}\nabla \cdot (\rho_h^{k+1}\mathbf{u}_h^k)\mathbf{u}_h^{k+1}, \mathbf{u}_h^{k+1} \right\rangle = 0.$$

Thus, we obtain

$$\|\sigma_h^{k+1}\mathbf{u}_h^{k+1}\|_{\mathbf{L}^2}^2 - \|\sigma_h^k \mathbf{u}_h^k\|_{\mathbf{L}^2}^2 + \|\sigma_h^k \delta \mathbf{u}_h^{k+1}\|_{\mathbf{L}^2}^2 + 2\mu\tau \|\nabla \mathbf{u}_h^{k+1}\|_{\mathbf{L}^2}^2 + 2\tau \langle \nabla p_h^k, \mathbf{u}_h^{k+1} \rangle = 0. \quad (4.27)$$

Since we are analyzing the non-incremental method, $\gamma = 0$ and $\phi_h^\flat = p_h^{k+1}$. Apply the operator δ to (4.24) and set $r_h = \delta p_h^{k+1}$ in the result. The Cauchy-Schwarz inequality and Hypothesis (4.20), imply that

$$\frac{\tau^2}{\chi} \|\nabla \delta p_h^{k+1}\|_{\mathbf{L}^2}^2 \leq \|\sigma_h^k \delta \mathbf{u}_h^{k+1}\|_{\mathbf{L}^2}^2 \quad (4.28)$$

Setting $r_h = 2\tau^2 p_h^k$ in (4.24), we derive

$$2\tau \langle \mathbf{u}_h^{k+1}, \nabla p_h^k \rangle = \tfrac{2\tau^2}{\chi} \langle \nabla p_h^{k+1}, \nabla p_h^k \rangle = \tfrac{\tau^2}{\chi} \left[\|\nabla p_h^{k+1}\|_{\mathbf{L}^2}^2 + \|\nabla p_h^k\|_{\mathbf{L}^2}^2 - \|\nabla \delta p_h^{k+1}\|_{\mathbf{L}^2}^2 \right]. \quad (4.29)$$

Adding (4.27) and (4.29), and using (4.28), we obtain

$$\|\sigma_h^{k+1}\mathbf{u}_h^{k+1}\|_{\mathbf{L}^2}^2 + 2\mu\tau\|\nabla\mathbf{u}_h^{k+1}\|_{\mathbf{L}^2}^2 + \frac{\tau^2}{\chi}\|\nabla p_h^{k+1}\|_{\mathbf{L}^2} \leq \|\sigma_h^k\mathbf{u}_h^k\|_{\mathbf{L}^2}^2$$

which when we add up over $k = 0, \ldots, K-1$ gives the desired stability result. □

Remark 28. The quantity $\frac{1}{2}\|\sigma_h^k\mathbf{u}_h^k\|_{\mathbf{L}^2}^2$ is the kinetic energy of the flow. Hence it is more natural to establish bounds in terms of this quantity than on the velocity itself; see also Lions [61].

Let us now prove stability estimates for the incremental scheme.

Theorem 11. *Assume that* (4.20) *holds. Then, for any* $\tau > 0$ *the solution* $\mathbf{u}_{h,\tau} \subset \mathbf{X}_h$ *and* $p_{h,\tau} \subset M_h$ *to the scheme of Section B with* $\gamma = 1$ *satisfies the following inequality:*

$$\|\sigma_h^K\mathbf{u}_h^K\|_{\mathbf{L}^2}^2 + 2\mu\tau\sum_{k=1}^{K}\|\nabla\mathbf{u}_h^k\|_{\mathbf{L}^2}^2 + \frac{\tau^2}{\chi}\|\nabla p_h^K\|_{\mathbf{L}^2}^2 + \frac{\tau^2}{\chi}\sum_{k=1}^{K-1}\|\nabla\delta p_h^k\|_{\mathbf{L}^2}^2 \leq \|\sigma_h^0\mathbf{u}_h^0\|_{\mathbf{L}^2}^2 + \frac{\tau^2}{\chi}\|\nabla p_h^0\|_{\mathbf{L}^2}^2,$$

where $\sigma_h^k = \sqrt{\rho_h^k}$.

Proof. In this case, $\gamma = 1$ and $\phi_h^\flat = \delta p_h^{k+1}$. Proceeding as in the proof of Theorem 10 we obtain the similar to (4.27) identity

$$\|\sigma_h^{k+1}\mathbf{u}_h^{k+1}\|_{\mathbf{L}^2}^2 - \|\sigma_h^k\mathbf{u}_h^k\|_{\mathbf{L}^2}^2 + \|\sigma_h^k\delta\mathbf{u}_h^{k+1}\|_{\mathbf{L}^2}^2 + 2\mu\tau\|\nabla\mathbf{u}_h^{k+1}\|_{\mathbf{L}^2}^2 + 2\tau\langle\nabla p_h^\sharp, \mathbf{u}_h^{k+1}\rangle = 0. \quad (4.30)$$

By (4.22), we infer

$$\begin{aligned} -2\tau\langle\nabla p_h^\sharp, \mathbf{u}_h^{k+1}\rangle &= -2\tau\langle\nabla(2p_h^k - p_h^{k-1}), \mathbf{u}_h^{k+1}\rangle \\ &= 2\tau\langle\nabla\delta^2 p_h^{k+1}, \mathbf{u}_h^{k+1}\rangle - 2\tau\langle\nabla p_h^{k+1}, \mathbf{u}_h^{k+1}\rangle. \end{aligned} \quad (4.31)$$

Now, in (4.24), set $r_h = \frac{2\tau}{\chi}\delta^2 p_h^{k+1}$. We obtain

$$-\frac{2\tau^2}{\chi}\langle\nabla\delta p_h^{k+1}, \nabla\delta p_h^{k+1} - \nabla\delta p_h^k\rangle + 2\tau\langle\mathbf{u}_h^{k+1}, \nabla\delta^2 p_h^{k+1}\rangle = 0.$$

Using the identity $2a\cdot(a-b) = a^2 - b^2 + (a-b)^2$ we obtain

$$\frac{\tau^2}{\chi}\left[-\|\nabla\delta p_h^{k+1}\|_{\mathbf{L}^2}^2 + \|\nabla\delta p_h^k\|_{\mathbf{L}^2}^2 - \|\nabla\delta^2 p_h^{k+1}\|_{\mathbf{L}^2}^2\right] + 2\tau\langle\mathbf{u}_h^{k+1}, \nabla\delta^2 p_h^{k+1}\rangle = 0. \quad (4.32)$$

Set $r_h = \frac{2\tau^2}{\chi}p_h^{k+1}$ in (4.24). We get

$$\frac{\tau^2}{\chi}\left[\|\nabla p_h^{k+1}\|_{\mathbf{L}^2}^2 - \|\nabla p_h^k\|_{\mathbf{L}^2}^2 + \|\nabla\delta p_h^{k+1}\|_{\mathbf{L}^2}^2\right] = 2\tau\langle\mathbf{u}_h^{k+1}, \nabla p_h^{k+1}\rangle, \quad (4.33)$$

where we used the identity mentioned before. Finally, apply the operator δ to (4.24). Using the lower bound Hypothesis (4.20), we derive the following estimate

$$\frac{\tau^2}{\chi}\|\nabla\delta^2 p_h^{k+1}\|_{\mathbf{L}^2}^2 \le \chi\|\delta\mathbf{u}_h^{k+1}\|_{\mathbf{L}^2}^2 \le \|\sigma_h^k\delta\mathbf{u}_h^{k+1}\|_{\mathbf{L}^2}^2. \quad (4.34)$$

Adding (4.30), (4.31), (4.32), (4.33) and (4.34), we obtain

$$\|\sigma_h^{k+1}\mathbf{u}_h^{k+1}\|_{\mathbf{L}^2}^2 - \|\sigma_h^k\mathbf{u}_h^k\|_{\mathbf{L}^2}^2 + 2\mu\tau\|\nabla\mathbf{u}_h^{k+1}\|_{\mathbf{L}^2}^2 + \frac{\tau^2}{\chi}\left[\|\nabla p_h^{k+1}\|_{\mathbf{L}^2}^2 - \|\nabla p_h^k\|_{\mathbf{L}^2}^2 + \|\nabla\delta p_h^k\|_{\mathbf{L}^2}^2\right] \le 0.$$

The desired result is obtained by adding up these relations for $n = 0, \ldots, N-1$. □

Remark 29. The above algorithm is an improvement over the second-order algorithm described [69, Algorithm 2], which requires a very strong (somewhat unrealistic) compatibility condition between the density and velocity spaces.

Remark 30. As usual for fractional time stepping techniques for the Stokes and Navier-Stokes equations, the stability property from Theorems 10 and 11 does not explicitly require the pair of spaces $(\mathbf{X}_h, M_h)$ to satisfy the LBB condition. This impression is misleading, since the estimates given by these Theorems do not give a

realistic stability on the pressure (unless $\tau \geq ch$). When going through the details one eventually realizes that the LBB condition must be invoked to prove stability on the pressure in $L^2(\Omega)$, we refer the reader to e.g. [38, 39, 44] for more details on this issue.

D. Error Estimates for the First-Order Scheme

The purpose of this section is to obtain error estimates for the algorithm (4.21)–(4.25). In order to do so, we must assume that the pair of spaces $(\mathbf{X}_h, M_h)$ satisfies a discrete inf–sup condition (cf. [34, 27]), i.e., there is $c > 0$ independent of h such that

$$\inf_{q_h \in M_h} \sup_{\mathbf{v}_h \in \mathbf{X}_h} \frac{\int_\Omega \mathbf{v}_h \cdot \nabla q_h}{\|q_h\|_{L^2} \|\mathbf{v}_h\|_{\mathbf{H}^1}} \geq c.$$

Moreover, we assume that the following approximation properties hold (cf. [34, 27]): There is $l \in \mathbb{N}$ such that for all $\ell \in [0, l]$

$$\inf_{r_h \in W_h} \|r - r_h\|_{L^2} \leq c h^{\ell+1} \|r\|_{H^{\ell+1}}, \quad \forall r \in H^{\ell+1}(\Omega). \tag{4.35}$$

$$\inf_{\mathbf{v}_h \in \mathbf{X}_h} \{\|v - \mathbf{v}_h\|_{\mathbf{L}^2} + h\|v - \mathbf{v}_h\|_{\mathbf{H}^1}\} \leq c h^{\ell+1} \|v\|_{\mathbf{H}^{\ell+1}}, \quad \forall v \in \mathbf{H}^{\ell+1}(\Omega) \cap \mathbf{H}_0^1(\Omega), \tag{4.36}$$

$$\inf_{q_h \in M_h} \|q - q_h\|_{L^2} \leq c h^{\ell} \|q\|_{H^{\ell}}, \qquad \forall q \in H^{\ell}(\Omega) \cap L_0^2(\Omega). \tag{4.37}$$

Remark 31. The references cited above provide several examples of spaces with these properties. A simple example is the following. Let $\mathcal{T}_h$ be a regular triangulation of $\bar{\Omega}$ composed of triangles in two dimensions (tetrahedra in three dimensions). Then, for

any $l \geq 1$ the spaces

$$W_h = \left\{ r_h \in \mathcal{C}^0(\bar{\Omega}) : \ r_h|_T \in \mathbb{P}_l, \ \forall T \in \mathcal{T}_h \right\},$$
$$\mathbf{X}_h = \left\{ \mathbf{v}_h \in \boldsymbol{\mathcal{C}}^0(\bar{\Omega}) : \ \mathbf{v}_h|_T \in \mathbb{P}_{l+1}, \ \forall T \in \mathcal{T}_h \right\},$$
$$M_h = \left\{ q_h \in \mathcal{C}^0(\bar{\Omega}) : \ q_h|_T \in \mathbb{P}_l, \ \forall T \in \mathcal{T}_h \right\},$$

satisfy all the hypotheses given above. If the triangulation consists of quadrilaterals (rectangular prisms) the same definitions with the polynomial space $\mathbb{P}$ replaced by $\mathbb{Q}$ also satisfy the hypotheses.

For any t in $[0,T]$ we define the Stokes projection of the solution $(\mathsf{u}(t), \mathsf{p}(t))$ of (1.3)–(1.4) as the pair $(\mathbf{w}_h(t), q_h(t)) \in \mathbf{X}_h \times M_h$ that solves

$$\begin{cases} \langle \nabla \mathbf{w}_h(t), \nabla \mathbf{v}_h \rangle + \langle \nabla q_h(t), \mathbf{v}_h \rangle = \langle \nabla \mathsf{u}(t), \nabla \mathbf{v}_h \rangle - \langle \mathsf{p}(t), \nabla \cdot \mathbf{v}_h \rangle, & \forall \mathbf{v}_h \in \mathbf{X}_h, \\ \langle \mathbf{w}_h(t), \nabla r_h \rangle = 0, & \forall r_h \in M_h. \end{cases} \tag{4.38}$$

Owing to the regularization properties of the Stokes operator, the following estimates hold:

Lemma 5. *If* $\mathsf{u} \in L^\beta\left(\mathbf{H}^{l+1}(\Omega) \cap \mathbf{H}_0^1(\Omega)\right)$ *and* $\mathsf{p} \in L^\beta\left(H^l(\Omega)\right)$ *for* $1 \leq \beta \leq \infty$*, then there exists* $c > 0$ *such that*

$$\|\mathsf{u} - \mathbf{w}_h\|_{L^\beta(\mathbf{L}^2)} + h\left[\|\mathsf{u} - \mathbf{w}_h\|_{L^\beta(\mathbf{H}^1)} + \|\mathsf{p} - q_h\|_{L^\beta(L^2)}\right] \leq c h^{l+1}\left[\|\mathsf{u}\|_{L^\beta(\mathbf{H}^{l+1})} + \|\mathsf{p}\|_{L^\beta(H^l)}\right]. \tag{4.39}$$

Moreover, if $\mathsf{u} \in L^\beta\left(\mathbf{H}^2(\Omega) \cap \mathbf{H}_0^1(\Omega)\right)$ *and* $\mathsf{p} \in L^\beta\left(H^1(\Omega)\right)$

$$\|\mathbf{w}_h\|_{L^\beta(\mathbf{L}^\infty \cap \mathbf{W}^{1,3})} + \|q_h\|_{L^\beta(H^1)} \leq c\left[\|\mathsf{u}\|_{L^\beta(\mathbf{H}^2)} + \|\mathsf{p}\|_{L^\beta(H^1)}\right]. \tag{4.40}$$

Concerning the initial approximations obtained in the Initialization step, we must

assume that

$$\|\rho_0 - \rho_h^0\|_{L^\infty} + \|\mathbf{u}_0 - \mathbf{u}_h^0\|_{\mathbf{L}^2} + h\|\mathbf{u}_0 - \mathbf{u}_h^0\|_{\mathbf{H}^1} + h\|p_0 - p_h^0\|_{L^2} \leq ch^{l+1}. \tag{4.41}$$

We begin by carrying out a consistency analysis of the schemes. To simplify the notation, we introduce the following functions to represent the errors:

$$\begin{cases} \eta(t) := \mathsf{u}(t) - \mathbf{w}_h(t), & \mu(t) := \mathsf{p}(t) - q_h(t), \\ \mathbf{e}_h^k := \mathbf{w}_h^k - \mathbf{u}_h^k, & \epsilon_h^k := q_h^k - p_h^k, \end{cases} \tag{4.42}$$

The functions $\eta(t)$ and $\mu(t)$ can be regarded as the interpolation errors, whereas the functions $\mathbf{e}_h^k$ and ϵ_h^k represent the approximation errors. In addition to (4.41), we make the following regularity assumptions on the exact solution of problem (1.3):

$$\rho \in W^{1,\infty}\left(W^{1,\infty}(\Omega)\right), \quad \mathsf{u} \in W^{1,\infty}\left(\mathbf{H}_0^1(\Omega) \cap \mathbf{H}^{l+1}(\Omega)\right), \quad \mathsf{p} \in W^{1,\infty}\left(H^l(\Omega)\right). \tag{4.43}$$

Let us now determine the equations that control the errors. By taking the difference between the first equation of (4.38) and (4.23) we obtain the equation that controls $\mathbf{e}_h^k$:

$$\left\langle \frac{\rho_h^\star \mathbf{e}_h^{k+1} - \rho_h^k \mathbf{e}_h^k}{\tau}, \mathbf{v}_h \right\rangle + \mu \left\langle \nabla \mathbf{e}_h^{k+1}, \nabla \mathbf{v}_h \right\rangle + \left\langle \nabla \left(q_h^{k+1} - p_h^\sharp \right), \mathbf{v}_h \right\rangle = \mathcal{R}^{k+1}(\mathbf{v}_h), \quad \forall \mathbf{v}_h \in \mathbf{X}_h, \tag{4.44}$$

where the residual $\mathcal{R}^{k+1}(\mathbf{v}_h)$ is decomposed as follows

$$\mathcal{R}^{k+1}(\mathbf{v}_h) = R_0^{k+1}(\mathbf{v}_h) + R_1^{k+1}(\mathbf{v}_h) + R_{nl}^{k+1}(\mathbf{v}_h), \tag{4.45}$$

and

$$R_0^{k+1}(\mathbf{v}_h) := \Big\langle \rho_h^k \frac{\mathbf{w}_h^{k+1} - \mathbf{w}_h^k}{\tau} - \rho^{k+1}\mathsf{u}_t^{k+1}, \mathbf{v}_h \Big\rangle, \tag{4.46}$$

$$R_1^{k+1}(\mathbf{v}_h) := \frac{1}{2}\Big\langle \frac{\rho_h^{k+1} - \rho_h^k}{\tau}\mathbf{w}_h^{k+1} - \rho_t^{k+1}\mathsf{u}^{k+1}, \mathbf{v}_h \Big\rangle, \tag{4.47}$$

$$R_{nl}^{k+1}(\mathbf{v}_h) := \left\langle \rho_h^{k+1}\mathbf{u}_h^k \cdot \nabla \mathbf{u}_h^{k+1} - \rho^{k+1}\mathsf{u}^{k+1}\cdot\nabla \mathsf{u}^{k+1}, \mathbf{v}_h \right\rangle \tag{4.48}$$

$$+ \frac{1}{2}\left\langle \nabla\cdot(\rho_h^{k+1}\mathbf{u}_h^k)\mathbf{u}_h^{k+1} - \nabla\cdot(\rho^{k+1}\mathsf{u}^{k+1})\mathsf{u}^{k+1}, \mathbf{v}_h \right\rangle. \tag{4.49}$$

To obtain the equation that controls the quantity ϵ_h^k we use (4.24) along with the property that $\langle \mathbf{w}_h, \nabla r_h\rangle = 0$ for all $r_h \in M_h$,

$$\left\langle \nabla \epsilon_h^\flat, \nabla r_h \right\rangle = \frac{\chi}{\tau}\left\langle \mathbf{e}_h^{k+1}, \nabla r_h \right\rangle + \left\langle \nabla q_h^\flat, \nabla r_h \right\rangle, \tag{4.50}$$

where for any sequence ψ_τ we henceforth denote

$$\psi^\flat = \psi^{k+1} - \gamma\psi^k, \qquad \text{and} \qquad \psi^\sharp = \psi^k + \delta\psi^k. \tag{4.51}$$

The two equations (4.44)–(4.50) will be used repeatedly in the error analysis.

The error analysis is based on energy arguments similar to those used to obtain stability in Section C. The first of these arguments consists of testing (4.44) with $\mathbf{v}_h := 2\tau\mathbf{e}_h^{k+1}$. Then, as in the proof of Theorem 10,

$$2\mathbf{e}_h^{k+1}\cdot(\rho_h^\star\mathbf{e}_h^{k+1} - \rho_h^k\mathbf{e}_h^k) = \rho_h^{k+1}|\mathbf{e}_h^{k+1}|^2 + \rho_h^k|\delta\mathbf{e}_h^{k+1}|^2 - \rho_h^k|\mathbf{e}_h^k|^2.$$

Testing (4.44) with $\mathbf{v}_h := 2\tau\mathbf{e}_h^{k+1}$ gives

$$\|\sigma_h^{k+1}\mathbf{e}_h^{k+1}\|_{\mathbf{L}^2}^2 - \|\sigma_h^k\mathbf{e}_h^k\|_{\mathbf{L}^2}^2 + \|\sigma_h^k\delta\mathbf{e}_h^{k+1}\|_{\mathbf{L}^2}^2 + 2\mu\tau\left\|\mathbf{e}_h^{k+1}\right\|_{\mathbf{H}^1}^2$$
$$+ 2\tau\left\langle \nabla\epsilon_h^\sharp, \mathbf{e}_h^{k+1}\right\rangle = 2\tau\left\langle \nabla(q_h^\sharp - q_h^{k+1}), \mathbf{e}_h^{k+1}\right\rangle + 2\tau\mathcal{R}^{k+1}(\mathbf{e}_h^{k+1}), \tag{4.52}$$

where, as before, $\sigma_h := \sqrt{\rho_h}$.

We finish the consistency analysis by giving an estimate on the consistency resid-

ual $2\tau\mathcal{R}^{k+1}(\mathbf{e}_h^{k+1})$. The following Lemma provides this estimate.

Lemma 6. *Assume that the solution to* (1.3)-(1.4) *satisfies* (4.43) *and that the sequence of approximate densities* $\{\rho_h^k\}$ *satisfies* (4.20)*. Then*

$$|\mathcal{R}^{k+1}(\mathbf{e}_h^{k+1})| \leq c\left[\tau + h^{l+1} + \|\rho_h^k - \rho^k\|_{L^2} + \left\|\frac{1}{\tau}\delta\rho_h^{k+1} - \rho_t^{k+1}\right\|_{L^2} + \|\rho_h^{k+1} - \rho^{k+1}\|_{H^1}\right]^2 + \frac{1}{2}\mu\|\mathbf{e}_h^{k+1}\|_{\mathbf{H}^1}^2 + c\|\sigma_h^k\mathbf{e}_h^k\|_{\mathbf{L}^2}^2. \quad (4.53)$$

Proof. We estimate separately each of the terms that compose $\mathcal{R}^{k+1}(\mathbf{e}_h^{k+1})$. For the first term

$$\begin{aligned}
R_0^{k+1}(\mathbf{e}_h^{k+1}) &= \left\langle \rho_h^k \frac{1}{\tau}\delta\mathbf{w}_h^{k+1} - \rho^{k+1}\mathsf{u}_t^{k+1}, \mathbf{e}_h^{k+1}\right\rangle \\
&= \left\langle \rho_h^k\left(\frac{1}{\tau}\delta\mathbf{w}_h^{k+1} - \mathsf{u}_t^{k+1}\right), \mathbf{e}_h^{k+1}\right\rangle + \left\langle (\rho_h^k - \rho^k)\mathsf{u}_t^{k+1}, \mathbf{e}_h^{k+1}\right\rangle \\
&\quad - \left\langle \delta\rho^{k+1}\mathsf{u}_t^{k+1}, \mathbf{e}_h^{k+1}\right\rangle \\
&\leq c\|\mathbf{e}_h^{k+1}\|_{\mathbf{L}^6}\left(\|\rho_h^k\|_{L^\infty}\left\|\frac{1}{\tau}\delta\mathbf{w}_h^{k+1} - \mathsf{u}_t^{k+1}\right\|_{\mathbf{L}^2}\right. \\
&\quad \left. + (\|\rho_h^k - \rho^k\|_{L^2} + \|\delta\rho^{k+1}\|_{L^2})\|u_t^{k+1}\|_{\mathbf{L}^3}\right) \\
&\leq c\|\mathbf{e}_h^{k+1}\|_{\mathbf{H}^1}\left(\tau + h^{l+1} + \|\rho_h^k - \rho^k\|_{L^2}\right),
\end{aligned}$$

where we used (4.39), (4.20), and (4.43) to derive the last inequality.

We proceed similarly for the second term,

$$\begin{aligned}
R_1^{k+1}(\mathbf{e}_h^{k+1}) &= \frac{1}{2}\left\langle \frac{1}{\tau}\delta\rho_h^{k+1}\mathbf{w}_h^{k+1} - \rho_t^{k+1}\mathsf{u}^{k+1}, \mathbf{e}_h^{k+1}\right\rangle \\
&= \frac{1}{2}\left\langle \left(\frac{1}{\tau}\delta\rho_h^{k+1} - \rho_t^{k+1}\right)\mathbf{w}_h^{k+1}, \mathbf{e}_h^{k+1}\right\rangle + \frac{1}{2}\left\langle \rho_t^{k+1}(\mathbf{w}_h^{k+1} - \mathsf{u}^{k+1}), \mathbf{e}_h^{k+1}\right\rangle \\
&\leq c\|\mathbf{e}_h^{k+1}\|_{\mathbf{L}^6}\left(\left\|\frac{1}{\tau}\delta\rho_h^{k+1} - \rho_t^{k+1}\right\|_{L^2}\|\mathbf{w}_h^{k+1}\|_{\mathbf{L}^3} + \|\rho_t^{k+1}\|_{L^3}\|\mathbf{w}_h^{k+1} - \mathsf{u}^{k+1}\|_{\mathbf{L}^2}\right) \\
&\leq c\|\mathbf{e}_h^{k+1}\|_{\mathbf{H}^1}\left(h^{l+1} + \left\|\frac{1}{\tau}\delta\rho_h^{k+1} - \rho_t^{k+1}\right\|_{L^2}\right),
\end{aligned}$$

where we used (4.39), (4.40), and (4.43) to derive the last inequality.

The derivation of an estimate for the nonlinear advection component of the residual is done by repeating an argument from [39]; we slightly modify the argument though to account for the fact that the density is not constant. We begin by noticing that, for functions that are smooth enough for the integrals to make sense, the following identity holds:

$$\langle \rho \mathbf{u}\cdot\nabla\mathbf{v}, \mathbf{v}\rangle + \frac{1}{2}\langle \nabla\cdot(\rho\mathbf{u})\mathbf{v}, \mathbf{v}\rangle = 0.$$

Then, using the above identity with $\mathbf{v} = \mathbf{e}_h$, we rewrite the term $R_{nl}^{k+1}(\mathbf{e}_h^{k+1})$ as follows

$$\begin{aligned} R_{nl}^{k+1}(\mathbf{e}_h^{k+1}) = &- \left\langle \rho_h^{k+1}\mathbf{e}_h^k\cdot\nabla\mathbf{w}_h^{k+1} + \tfrac{1}{2}\nabla\cdot(\rho_h^{k+1}\mathbf{e}_h^k)\mathbf{w}_h^{k+1}, \mathbf{e}_h^{k+1}\right\rangle \\ &+ \left\langle (\rho_h^{k+1}-\rho^{k+1})\mathbf{w}_h^k\cdot\nabla\mathbf{w}_h^{k+1} + \frac{1}{2}\nabla\cdot((\rho_h^{k+1}-\rho^{k+1})\mathbf{w}_h^k)\mathbf{w}_h^{k+1}, \mathbf{e}_h^{k+1}\right\rangle \\ &+ \left\langle \rho^{k+1}\left(\mathbf{w}_h^k\cdot\nabla\mathbf{w}_h^{k+1} - \mathsf{u}^{k+1}\cdot\nabla\mathsf{u}^{k+1}\right)\right. \\ &\left. + \frac{1}{2}\left(\nabla\cdot(\rho^{k+1}\mathbf{w}_h^k)\mathbf{w}_h^{k+1} - \nabla\cdot(\rho^{k+1}\mathsf{u}^{k+1})\mathsf{u}^{k+1}\right), \mathbf{e}_h^{k+1}\right\rangle \\ &:= A_1 + A_2 + A_3 \end{aligned}$$

Since the approximate density sequence $\{\rho_h^k\}$ satisfies (4.20) and the approximate velocity sequence $\{\mathbf{w}_h^k\}$ satisfies (4.40), we infer

$$A_1 \le c\|\sigma_h^k\mathbf{e}_h^k\|_{\mathbf{L}^2}\|\mathbf{e}_h^{k+1}\|_{\mathbf{H}^1},$$

where we estimated the second term after integrating it by parts, which is possible given the smoothness of $\mathbf{w}_h^{k+1}$ and $\mathbf{e}_h^{k+1}$. Using (4.40) we obtain

$$A_2 \le c\|\rho_h^{k+1}-\rho^{k+1}\|_{H^1}\|\mathbf{e}_h^{k+1}\|_{\mathbf{H}^1},$$

where, again, we integrated by parts the second term. Finally, given the smoothness of ρ^{k+1}, an estimate of A_3 is obtained by proceeding as in the constant density case,

see e.g. [39, 55]:

$$A_3 \leq c(\tau + h^{l+1})\|\mathbf{e}_h^{k+1}\|_{\mathbf{H}^1}.$$

The estimate (4.53) is obtained by combining the results above. □

As stated the results of Section C show, the stability of the algorithm that we are analyzing only marginally depends on the method which is used to approximate the density; the only assumption we make to achieve stability is that the algorithm that solves the mass conservation equation satisfies (4.20). Of course (4.20) is not sufficient to obtain error estimates. Performing the full error analysis would require to analyze the nonlinear coupling between the mass conservation equation and the momentum conservation equation. This would require to be specific on the type of approximation which is used to compute the approximate density field and would probably lead to lengthy technicalities of little interest. We are not going to do the full convergence analysis to avoid technicalities and to remain as general as possible on the way the mass conservation equation is approximated. We assume instead that, in some way, we are capable of computing an approximate density sequence $\{\rho_h^k\} \subset W_h$ from the knowledge of the approximated velocity sequence $\{\mathbf{u}_h^k\} \subset \mathbf{X}_h$. To be more specific we assume that the following holds:

$$\|(\rho - \rho_h)_\tau\|^2_{\ell^\infty(H^1)} + \left\| \left(\rho_t - \frac{\delta \rho_h}{\tau} \right)_\tau \right\|^2_{\ell^\infty(L^2)} \leq c(\lambda)(\tau + h^{l+1})^2 + \lambda \|\mathbf{e}_h^{k+1}\|^2_{\mathbf{H}^1} + c(\lambda)\|\sigma_h^k \mathbf{e}_h^k\|^2_{\mathbf{L}^2}, \quad (4.54)$$

where $\lambda \geq 0$ can be chosen as small as necessary. Given this assumption, the residual term $\mathcal{R}(\mathbf{e}_h^{k+1})$ simplifies as follows:

Corollary 5. *Assume that* (4.54) *holds. Then, the following estimate holds under*

the regularity assumptions of Lemma 6:

$$2\tau|\mathcal{R}^{k+1}(\mathbf{e}_h^{k+1})| \leq c\tau(\tau + h^{l+1})^2 + \mu\tau\|\mathbf{e}_h^{k+1}\|_{\mathbf{H}^1}^2 + c\tau\|\sigma^k\mathbf{e}_h^k\|_{\mathbf{L}^2}^2. \tag{4.55}$$

Proof. Use (4.53) where all the terms that involve differences of ρ_h and ρ can be majored by (4.54). The parameter λ is chosen so that $\lambda = \varepsilon\mu$, where ε is chosen small enough. □

We now consider the non-incremental and the incremental versions the algorithm separately.

As we have stated before, the non-incremental version of the method is obtained by setting $\gamma = 0$. Under assumption (4.54), the main error estimate for this algorithm is the following.

Theorem 12. *Assume that the solution to* (1.3)–(1.4) *satisfies* (4.43)*, and that* (4.20) *hold for all* $0 \leq k \leq K$*. Let* $(\mathbf{u}_h)_\tau$ *be the solution of* (4.23)–(4.24) *with* $\gamma = 0$ *and assume that* (4.41) *and* (4.54) *hold. Then*

$$\|\mathsf{u}_\tau - (\mathbf{u}_h)_\tau\|_{\ell^\infty(\mathbf{L}^2)} \leq c\left(h^{l+1} + \tau^{1/2}\right), \qquad \|\mathsf{u}_\tau - (\mathbf{u}_h)_\tau\|_{\ell^2(\mathbf{H}^1)} \leq c\left(h^l + \tau^{1/2}\right). \tag{4.56}$$

Conjecture 1. We expect that further regularity assumptions combined with a standard duality argument, e.g. multiplying the error equation by $S\mathbf{e}_h^{k+1}$, where S is the solution operator to the time-independent Stokes problem, should allow us to conclude that the following estimate holds in addition to (4.56):

$$\|\mathsf{u}_\tau - (\mathbf{u}_h)_\tau\|_{\ell^2(\mathbf{L}^2)} \leq c\left(h^{l+1} + \tau\right).$$

The reader is referred to [52, 39] for more details.

Remark 32. The error estimate (4.56) shows that, at least under assumption (4.54), the non-incremental fractional time-stepping technique for variable density fluid flows

performs as well as the analogous non-incremental pressure-correction scheme for constant density flows (see Theorem 7).

Proof. [Theorem 12] In this case $p_h^\sharp = p_h^k$ and $\phi_h^\flat = p_h^{k+1}$. Setting $r_h := 2\tau^2\epsilon_h^k/\chi$ in (4.50) we obtain

$$\frac{\tau^2}{\chi}\left[\left\|\nabla\epsilon_h^{k+1}\right\|_{\mathbf{L}^2}^2 + \left\|\nabla\epsilon_h^k\right\|_{\mathbf{L}^2}^2 - \left\|\nabla\delta\epsilon_h^{k+1}\right\|_{\mathbf{L}^2}^2\right] - 2\tau\left\langle\nabla\epsilon_h^k, \mathbf{e}_h^{k+1}\right\rangle = \frac{2\tau^2}{\chi}\left\langle\nabla q_h^{k+1}, \nabla\epsilon_h^k\right\rangle. \tag{4.57}$$

Next, apply δ to (4.50) and set $r_h := \tau\delta\epsilon_h^{k+1}$. The Cauchy-Schwarz inequality implies

$$\tau^2\left\|\nabla\delta\epsilon_h^{k+1}\right\|_{\mathbf{L}^2}^2 \le \left\|\chi\delta\mathbf{e}_h^{k+1} + \tau\nabla\delta q_h^{k+1}\right\|_{\mathbf{L}^2}^2 = \chi^2\left\|\delta\mathbf{e}_h^{k+1}\right\|_{\mathbf{L}^2}^2 + \tau^2\left\|\nabla\delta q_h^{k+1}\right\|_{\mathbf{L}^2}^2 + 2\chi\tau\left\langle\nabla\delta q_h^{k+1}, \delta\mathbf{e}_h^{k+1}\right\rangle,$$

which, by (4.20), implies

$$\frac{\tau^2}{\chi}\left\|\nabla\delta\epsilon_h^{k+1}\right\|_{\mathbf{L}^2}^2 \le \left\|\sigma_h^k\delta\mathbf{e}_h^{k+1}\right\|_{\mathbf{L}^2}^2 + \frac{\tau^2}{\chi}\left\|\nabla\delta q_h^{k+1}\right\|_{\mathbf{L}^2}^2 + 2\tau\left\langle\nabla\delta q_h^{k+1}, \delta\mathbf{e}_h^{k+1}\right\rangle. \tag{4.58}$$

Adding up (4.53), (4.57) and (4.58) and using Corollary 5, we obtain,

$$\|\sigma_h^{k+1}\mathbf{e}_h^{k+1}\|_{\mathbf{L}^2}^2 + \mu\tau\|\mathbf{e}_h^{k+1}\|_{\mathbf{H}^1}^2 + \frac{\tau^2}{\chi}\left[\|\nabla\epsilon_h^{k+1}\|_{\mathbf{L}^2}^2 + \|\nabla\epsilon_h^k\|_{\mathbf{L}^2}^2\right] \le$$
$$c\tau(\tau+h^{l+1})^2 + (1+c\tau)\|\sigma_h^k\mathbf{e}_h^k\|_{\mathbf{L}^2}^2 - 2\tau\left\langle\nabla\delta q_h^{k+1}, \mathbf{e}_h^k\right\rangle + \frac{\tau^2}{\chi}\|\delta q_h^{k+1}\|^2 + \frac{2\tau^2}{\chi}\left\langle\nabla q_h^{k+1}, \nabla\epsilon_h^k\right\rangle.$$

We estimate the last three terms in the right-hand side separately. Integrating by parts and using (4.40), the first one can be estimated as follows:

$$-2\tau\left\langle\nabla\delta q_h^{k+1}, \mathbf{e}_h^k\right\rangle \le 2\tau\|\delta q_h^{k+1}\|_{L^2}\|\mathbf{e}_h^k\|_{\mathbf{H}^1} \le c\tau^3 + \frac{\mu\tau}{2}\|\mathbf{e}_h^k\|_{\mathbf{H}^1}^2.$$

Similarly, the second term is estimated as follows:

$$\frac{\tau^2}{\chi}\|\nabla\delta q_h^{k+1}\|_{\mathbf{L}^2}^2 \le c\tau^3.$$

For the last term, using again (4.40) we obtain

$$\frac{2\tau^2}{\chi}\left\langle \nabla q_h^{k+1}, \nabla \epsilon_h^k \right\rangle \leq c\frac{\tau^2}{\chi}\|\nabla \epsilon_h^k\|_{\mathbf{L}^2} \leq c\tau^2 + \frac{\tau^2}{\chi}\|\nabla \epsilon_h^k\|_{\mathbf{L}^2}^2.$$

Notice that this term is responsible for the loss of optimality, i.e., full first-order accuracy is lost at this point.

Combining the above observations, we finally obtain

$$\|\sigma_h^{k+1}\mathbf{e}_h^{k+1}\|_{\mathbf{L}^2}^2 + \mu\tau\|\mathbf{e}_h^{k+1}\|_{\mathbf{H}^1}^2 \leq (1+c\tau)\|\sigma_h^k\mathbf{e}_h^k\|_{\mathbf{L}^2}^2 + \frac{\mu\tau}{2}\|\mathbf{e}_h^k\|_{\mathbf{H}^1}^2 + c\tau(\tau^{1/2}+h^{l+1})^2,$$

which, by the discrete Grönwall lemma implies

$$\|(\sigma_h\mathbf{e}_h)_\tau\|_{\ell^\infty(\mathbf{L}^2)} + \|(\mathbf{e}_h)_\tau\|_{\ell^2(\mathbf{H}^1)} \leq c(\tau^{1/2}+h^{l+1}).$$

The claimed error estimates follow from the triangle inequality, the definition

$$\mathsf{u}^k - \mathbf{u}_h^k = \eta^k + \mathbf{e}_h^k,$$

and (in the case of the $\ell^\infty(\mathbf{L}^2)$-norm) assumption (4.20). Notice that it is only at this point that the interpolation error in the $\mathbf{H}^1$-norm, which is of order $\mathcal{O}(h^l)$, is introduced. This a well-known super-convergence effect induced by our particular choice for the pair $(\mathbf{w}_h, q_h)$, see (4.38) and [86]. □

The incremental version of the algorithm is obtained by setting $\gamma = 1$. Under assumption (4.54), the main error estimate for this algorithm is stated as follows.

Theorem 13. *Assume that the solution to* (1.3)–(1.4) *satisfies* (4.43)*, and that* (4.20) *hold for all* $0 \leq k \leq K$*. Let* $(\mathbf{u}_h)_\tau$ *be the solution of* (4.23)–(4.24) *with* $\gamma = 1$ *and*

assume that (4.41) *and* (4.54) *hold. Then*

$$\|\mathsf{u}_\tau - (\mathbf{u}_h)_\tau\|_{\ell^\infty(\mathbf{L}^2)} \leq c\left(\tau + h^{l+1}\right), \tag{4.59}$$

$$\|\mathsf{u}_\tau - (\mathbf{u}_h)_\tau\|_{\ell^2(\mathbf{H}^1)} \leq c\left(\tau + h^{l}\right). \tag{4.60}$$

Remark 33. The error estimates from Theorem 13 show that, under the given assumptions on the density approximation, the incremental pressure-correction algorithm for variable density fluid flows performs as well as the analogous incremental projection-type pressure-correction scheme for constant density flows (cf. [44]).

Proof. [Theorem 13] In this case $p_h^\sharp = 2p_h^k - p_h^{k-1}$ and $\phi_h^\flat = \delta p_h^{k+1}$. Setting $r_h := -2\tau^2\delta^2\epsilon_h^{k+1}/\chi$ in (4.50), we obtain

$$-\frac{\tau^2}{\chi}\left[\|\nabla\delta\epsilon_h^{k+1}\|_{\mathbf{L}^2}^2 - \|\nabla\delta\epsilon_h^{k}\|_{\mathbf{L}^2}^2 + \|\nabla\delta^2\epsilon_h^{k+1}\|_{\mathbf{L}^2}^2\right] + 2\tau\left\langle \mathbf{e}_h^{k+1}, \nabla\delta^2\epsilon_h^{k+1}\right\rangle = -\frac{2\tau^2}{\chi}\left\langle \nabla\delta q_h^{k+1}, \nabla\delta^2\epsilon_h^{k+1}\right\rangle.$$

Setting $r_h := 2\tau^2\epsilon_h^{k+1}/\chi$ in (4.50), we obtain

$$\frac{\tau^2}{\chi}\left[\|\nabla\epsilon_h^{k+1}\|_{\mathbf{L}^2}^2 - \|\nabla\epsilon_h^{k}\|_{\mathbf{L}^2}^2 + \|\nabla\delta\epsilon_h^{k+1}\|_{\mathbf{L}^2}^2\right] = 2\tau\left\langle \mathbf{e}_h^{k+1}, \nabla\epsilon_h^{k+1}\right\rangle + \frac{2\tau^2}{\chi}\left\langle \nabla\delta q_h^{k+1}, \nabla\epsilon_h^{k+1}\right\rangle$$

Adding these two equations we arrive at

$$\frac{\tau^2}{\chi}\left[\|\nabla\epsilon_h^{k+1}\|_{\mathbf{L}^2}^2 - \|\nabla\epsilon_h^{k}\|_{\mathbf{L}^2}^2 + \|\nabla\delta\epsilon_h^{k}\|_{\mathbf{L}^2}^2\right] - \frac{\tau^2}{\chi}\|\nabla\delta^2\epsilon_h^{k+1}\|_{\mathbf{L}^2}^2 - 2\tau\left\langle \mathbf{e}_h^{k+1}, \nabla\epsilon_h^{\sharp}\right\rangle = \frac{2\tau^2}{\chi}\left\langle \nabla\delta q_h^{k+1}, \nabla\epsilon_h^{\sharp}\right\rangle. \tag{4.61}$$

Now we apply δ to (4.50) and we set $r_h := \tau\delta^2\epsilon_h^{k+1}$. The Cauchy-Schwarz in-

equality implies

$$\tau^2\|\nabla\delta^2\epsilon_h^{k+1}\|^2_{\mathbf{L}^2} \le \left\|\chi\delta\mathbf{e}_h^{k+1} + \tau\nabla\delta^2 q_h^{k+1}\right\|^2_{\mathbf{L}^2} = \\ \chi^2\|\delta\mathbf{e}_h^{k+1}\|^2_{\mathbf{L}^2} + \tau^2\|\nabla\delta^2 q_h^{k+1}\|^2_{\mathbf{L}^2} + 2\tau\chi\left\langle\nabla\delta^2 q_h^{k+1}, \delta\mathbf{e}_h^{k+1}\right\rangle,$$

and owing to (4.20) we infer

$$\frac{\tau^2}{\chi}\|\nabla\delta^2\epsilon_h^{k+1}\|^2_{\mathbf{L}^2} \le \|\sigma^k\delta\mathbf{e}_h^{k+1}\|^2_{\mathbf{L}^2} + \frac{\tau^2}{\chi}\|\nabla\delta^2 q_h^{k+1}\|^2_{\mathbf{L}^2} + 2\tau\left\langle\nabla\delta^2 q_h^{k+1}, \delta\mathbf{e}_h^{k+1}\right\rangle. \qquad (4.62)$$

Adding (4.52), (4.61) and (4.62), and using Corollary 5, we arrive at

$$\|\sigma_h^{k+1}\mathbf{e}_h^{k+1}\|^2_{\mathbf{L}^2} + \mu\tau\|\mathbf{e}_h^{k+1}\|^2_{\mathbf{H}^1} + \frac{\tau^2}{\chi}[\|\nabla\epsilon_h^{k+1}\|^2_{\mathbf{L}^2} + \|\nabla\delta\epsilon_h^{k+1}\|^2_{\mathbf{L}^2}] \le \\ (1+c\tau)\|\sigma_h^k\mathbf{e}_h^k\|^2_{\mathbf{L}^2} + \frac{\tau^2}{\chi}\|\nabla\epsilon_h^k\|^2_{\mathbf{L}^2} \\ + c\tau(\tau+h^{l+1})^2 + \frac{\tau^2}{\chi}\|\nabla\delta^2 q_h^{k+1}\|^2_{\mathbf{L}^2} - 2\tau\left\langle\nabla\delta^2 q_h^{k+1}, \mathbf{e}_h^k\right\rangle + \frac{2\tau^2}{\chi}\left\langle\nabla\delta q_h^{k+1}, \nabla\epsilon_h^\sharp\right\rangle.$$

Let us estimate the last three terms separately. Clearly,

$$\tau^2/\chi\|\nabla\delta^2 q_h^{k+1}\|^2_{\mathbf{L}^2} \le c\tau^3.$$

The second term is bounded from above as follows:

$$-2\tau\left\langle\nabla\delta^2 q_h^{k+1}, \mathbf{e}_h^k\right\rangle \le c\tau^3 + \frac{\mu\tau}{2}\|\mathbf{e}_h^k\|^2_{\mathbf{H}^1}.$$

Finally, for the third term we have

$$\frac{2\tau^2}{\chi}\left\langle\nabla\delta q_h^{k+1}, \nabla\epsilon_h^\sharp\right\rangle \le c\tau^3 + \tau^3\|\nabla\epsilon_h^k\|^2_{\mathbf{L}^2} + \tau^3\|\nabla\delta\epsilon_h^k\|^2_{\mathbf{L}^2}.$$

We obtain the estimate (4.59)-(4.60) by finishing as in the proof of Theorem 12. □

E. A Second-Order Fractional Time-Stepping Method

We have established in the previous section that the incremental version of the scheme (4.20)–(4.25) is first-order accurate in time both for the $\mathbf{L}^2$- and the $\mathbf{H}^1$-norm of the velocity. However, as shown in [39], we expect that the splitting error of the algorithm is second-order since the pressure term

$$p_h^\sharp = 2p_h^k - p_h^{k-1}, \tag{4.63}$$

that appears in the approximate momentum equation is a second-order extrapolation of the pressure p_h^{k+1}. This observation is the main motivation for our introducing a variant of the incremental method using a second-order approximation of the time derivative of the velocity.

Keeping the same notation as in the previous sections, the second-order variant of the algorithm is composed of the following steps:

Initialization First, we choose a penalty parameter χ as in the Initialization step of Section B. Next, we define $(\rho_h^0, \mathbf{u}_h^0, p_h^0, \phi_h^0 = 0) \in W_h \times \mathbf{X}_h \times M_h \times M_h$ to be a suitable approximation of the initial data of the problem. Then we compute an approximation of the exact solution at time $t = \tau$, say $(\rho_h^1, \mathbf{u}_h^1, p_h^1, \phi_h^1 = p_h^1 - p_h^0) \in W_h \times \mathbf{X}_h \times M_h \times M_h$.

Time-Stepping Given $(\rho_h^k, \mathbf{u}_h^k, p_h^k, \phi_h^k) \in W_h \times \mathbf{X}_h \times M_h \times M_h$ for $1 \leq k \leq K-1$, we compute the next time-step approximation as follows:

Density Update We are not specific on the way $\rho_h^{k+1} \in W_h$ is computed, but we assume that (4.20) holds and that there is a uniform constant M so that

$$\max_{0 \leq k \leq K-1} \left\| \frac{\rho_h^{k+1} - \rho_h^k}{\tau} \right\|_{L^\infty} \leq M\chi. \tag{4.64}$$

Velocity Update Similarly to the Velocity Update step of Section B we define

$$\rho_h^\star := \frac{3}{2}\rho_h^{k+1} - \frac{2}{3}\rho_h^k + \frac{1}{6}\rho_h^{k-1} = \rho_h^{k+1} + \frac{1}{6}(3\rho_h^{k+1} - 4\rho_h^k + \rho_h^{k-1}), \tag{4.65}$$

$$p_h^\sharp := p_h^k + \frac{4}{3}\phi_h^k - \frac{1}{3}\phi_h^{k-1}. \tag{4.66}$$

Then we compute $\mathbf{u}_h^{k+1} \in \mathbf{X}_h$ so that the following holds:

$$\begin{aligned}\left\langle \frac{3\rho_h^\star \mathbf{u}_h^{k+1} - 4\rho_h^{k+1}\mathbf{u}_h^k + \rho_h^{k+1}\mathbf{u}_h^{k-1}}{2\tau}, \mathbf{v}_h \right\rangle \\ + \left\langle \rho_h^{k+1}\mathbf{u}_h^\star \cdot \nabla \mathbf{u}_h^{k+1} + \frac{1}{2}\mathbf{u}_h^{k+1}\nabla\cdot(\rho_h^{k+1}\mathbf{u}_h^\star), \mathbf{v}_h \right\rangle \\ + \mu \left\langle \nabla \mathbf{u}_h^{k+1}, \nabla \mathbf{v}_h \right\rangle + \left\langle \nabla p_h^\sharp, \mathbf{v}_h \right\rangle = \left\langle f^{k+1}, \mathbf{v}_h \right\rangle, \quad \forall \mathbf{v}_h \in \mathbf{X}_h,\end{aligned} \tag{4.67}$$

where

$$\mathbf{u}_h^\star := 2\mathbf{u}_h^k - \mathbf{u}_h^{k-1}, \tag{4.68}$$

is a second-order extrapolation of the velocity.

Penalty We compute the pressure correction $\phi_h^{k+1} \in M_h$ so that the following holds:

$$\left\langle \nabla \phi_h^{k+1}, \nabla r_h \right\rangle = \frac{3\chi}{2\tau} \left\langle \mathbf{u}_h^{k+1}, \nabla r_h \right\rangle, \quad \forall r_h \in M_h. \tag{4.69}$$

Pressure Update Finally, the pressure is updated by setting

$$p_h^{k+1} = p_h^k + \phi_h^{k+1}. \tag{4.70}$$

Remark 34. The quantities $(\rho_h^1, \mathbf{u}_h^1, p_h^1, \phi_h^1)$ can be computed by using one step of the incremental first-order scheme described in Section B.

Remark 35. The term $\left\langle \frac{1}{2}\nabla\cdot(\rho_h^{k+1}\mathbf{u}_h^\star)\mathbf{u}_h^{k+1}, \mathbf{v}_h \right\rangle$ has been added to the equation to obtain unconditional stability with respect to the advection term. As in the proof of Lemma 6

we are going to use the following identity:

$$\left\langle \rho_h^{k+1}\mathbf{u}_h^\star\cdot\nabla\mathbf{u}_h^{k+1}+\frac{1}{2}\nabla\cdot(\rho_h^{k+1}\mathbf{u}_h^\star)\mathbf{u}_h^{k+1},\mathbf{u}_h^{k+1}\right\rangle = \int_\Omega \rho_h^{k+1}\mathbf{u}_h^\star\cdot\nabla\mathbf{u}_h^{k+1}\cdot\mathbf{u}_h^{k+1}+\frac{1}{2}\int_\Omega\nabla\cdot(\rho_h^{k+1}\mathbf{u}_h^\star)|\mathbf{u}_h^{k+1}|^2=0.$$

Remark 36. The term

$$\frac{3\rho_h^\star\mathbf{u}_h^{k+1}-4\rho_h^{k+1}\mathbf{u}_h^k+\rho_h^{k+1}\mathbf{u}_h^{k-1}}{2\tau}+\frac{1}{2}\nabla\cdot(\rho_h^{k+1}\mathbf{u}_h^\star)\mathbf{u}_h^{k+1},$$

is a second-order approximation of $[\rho_h\mathbf{u}_{h,t}](t^{k+1})$. Indeed, if the involved functions are smooth enough in time, we infer from the definition of $\rho_h^\star$ that

$$\begin{aligned}&\frac{3\rho_h^\star\mathbf{u}_h^{k+1}-4\rho_h^{k+1}\mathbf{u}_h^k+\rho_h^{k+1}\mathbf{u}_h^{k-1}}{2\tau}+\frac{1}{2}\nabla\cdot(\rho_h^{k+1}\mathbf{u}_h^\star)\mathbf{u}_h^{k+1}=\\ &\quad\frac{\rho_h^{k+1}}{2\tau}(3\mathbf{u}_h^{k+1}-4\mathbf{u}_h^k+\mathbf{u}_h^{k-1})+\frac{1}{2}\left(\frac{3\rho_h^{k+1}-4\rho_h^k+\rho_h^{k-1}}{2\tau}+\nabla\cdot(\rho_h^{k+1}\mathbf{u}_h^\star)\right)\mathbf{u}_h^{k+1}=\\ &\qquad[\rho_h\mathbf{u}_{h,t}]^{k+1}+\frac{1}{2}\left[\rho_{h,t}+\nabla\cdot(\rho_h\mathbf{u}_h)\right]^{k+1}\mathbf{u}_h^{k+1}+\mathcal{O}(\tau^2)=[\rho_h\mathbf{u}_{h,t}]^{k+1}+\mathcal{O}(\tau^2),\end{aligned}$$

which proves the claim.

We now establish stability for the algorithm (4.67)-(4.69)-(4.70). Again, to avoid irrelevant technicalities, assume that $\mathbf{f}\equiv 0$. The stability of the scheme is given by the following Theorem.

Theorem 14. *Assume that the sequence of approximate densities* $\{\rho_h^k\}_{k\geq 0}\subset W_h$ *satisfies* (4.20) *and* (4.64). *Then, for* τ *small enough, the sequence* $\{(\mathbf{u}_h^k,p_h^k)\}_{k\geq 0}\subset \mathbf{X}_h\times M_h$ *obtained by the algorithm* (4.67)-(4.69)-(4.70) *satisfies the following estimate:*

$$\begin{aligned}&\|\sigma_h^k\mathbf{u}_h^k\|_{\mathbf{L}^2}^2+\mu\tau\|\mathbf{u}_h^k\|_{\mathbf{H}^1}^2+\frac{\tau^2}{\chi}\|\nabla p_h^k\|_{\mathbf{L}^2}^2+\frac{\tau^2}{\chi}\|\nabla\delta p_h^{k-1}\|^2\leq\\ &\qquad\mathcal{K}(1+e^{cT})\left(\|\sigma_h^0\mathbf{u}_h^0\|_{\mathbf{L}^2}^2+\|\sigma_h^1\mathbf{u}_h^1\|_{\mathbf{L}^2}^2+\|\nabla p_h^0\|_{\mathbf{L}^2}^2+\|\nabla p_h^1\|_{\mathbf{L}^2}^2\right),\quad\forall k\geq 2,\end{aligned}\tag{4.71}$$

for some constants c *and* $\mathcal{K}$.

Proof. Note first that, as already mentioned in Remark 36, the time derivative can be re-written as follows:

$$\frac{3\rho_h^\star \mathbf{u}_h^{k+1} - 4\rho_h^{k+1}\mathbf{u}_h^k + \rho_h^{k+1}\mathbf{u}_h^{k-1}}{2\tau} = \rho^{k+1}\frac{3\mathbf{u}_h^{k+1} - 4\mathbf{u}_h^k + \mathbf{u}_h^{k-1}}{2\tau} + \frac{1}{2}\mathbf{u}_h^{k+1}\frac{3\rho^{k+1} - 4\rho^k + \rho^{k-1}}{2\tau},$$

which is an approximation of $\rho\mathsf{u}_t + \frac{1}{2}\mathsf{u}\rho_t$. Once tested with u, the expression $(\rho\mathsf{u}_t + \frac{1}{2}\mathsf{u}\rho_t)\mathsf{u}$ gives $(\frac{1}{2}\rho\mathsf{u}^2)_t$, and after integration over Ω and over the time interval $(0,T)$ this yields kinetic energy conservation. We have been able to reproduce this argument at the discrete level for the first-order time stepping described in Section B, see (4.52). Unfortunately, we have not yet figured out how to repeat this argument with BDF2. We are going to content ourselves with a sub-optimal stability analysis which will yield the growth constant $(1 + e^{cT})$ in (4.71).

Using Assumption (4.64), we have the following estimate

$$\begin{aligned}
\left\langle \left(3\rho_h^{k+1} - 4\rho_h^k + \rho_h^{k-1}\right)\mathbf{u}_h^{k+1}, \mathbf{u}_h^{k+1}\right\rangle &= 3\int_\Omega \left(\rho_h^{k+1} - \rho_h^k\right)|\mathbf{u}_h^{k+1}|^2 \\
&\quad - \int_\Omega \left(\rho_h^k - \rho_h^{k-1}\right)|\mathbf{u}_h^{k+1}|^2 \\
&\geq -\left(3\left\|\frac{\rho_h^{k+1} - \rho_h^k}{\chi}\right\|_{L^\infty}\right. \\
&\quad \left. + \left\|\frac{\rho_h^k - \rho_h^{k-1}}{\chi}\right\|_{L^\infty}\right)\|\sigma_h^{k+1}\mathbf{u}_h^{k+1}\|_{\mathbf{L}^2}^2 \\
&\geq -4M\tau\|\sigma_h^{k+1}\mathbf{u}_h^{k+1}\|_{\mathbf{L}^2}^2.
\end{aligned}$$

A similar treatment gives

$$\begin{aligned}
2\left\langle \rho_h^{k+1}\left(3\mathbf{u}_h^{k+1} - 4\mathbf{u}_h^k + \mathbf{u}_h^{k-1}\right), \mathbf{u}_h^{k+1}\right\rangle &\geq 3\|\sigma_h^{k+1}\mathbf{u}_h^{k+1}\|_{\mathbf{L}^2}^2 - (4 + 8M\tau)\|\sigma_h^k\mathbf{u}_h^k\|_{\mathbf{L}^2}^2 \\
&\quad + (1 - 6M\tau)\|\sigma_h^{k-1}\mathbf{u}_h^{k-1}\|_{\mathbf{L}^2}^2 + 2\|\sigma_h^{k+1}\delta\mathbf{u}_h^{k+1}\|_{\mathbf{L}^2}^2 - 2\|\sigma_h^k\delta\mathbf{u}_h^k\|_{\mathbf{L}^2}^2 + \|\sigma_h^{k+1}\delta^2\mathbf{u}_h^{k+1}\|_{\mathbf{L}^2}^2.
\end{aligned}$$

Combining the above two inequalities gives

$$2\left\langle 3\rho_h^\star \mathbf{u}_h^{k+1} - 4\rho_h^{k+1}\mathbf{u}_h^k + \rho_h^{k+1}\mathbf{u}_h^{k-1}, \mathbf{u}_h^{k+1}\right\rangle \geq (3-4M\tau)\|\sigma_h^{k+1}\mathbf{u}_h^{k+1}\|_{\mathbf{L}^2}^2 - (4+8M\tau)\|\sigma_h^k\mathbf{u}_h^k\|_{\mathbf{L}^2}^2 + (1-6M\tau)\|\sigma_h^{k-1}\mathbf{u}_h^{k-1}\|_{\mathbf{L}^2}^2 + 2\|\sigma_h^{k+1}\delta\mathbf{u}_h^{k+1}\|_{\mathbf{L}^2}^2 - 2\|\sigma_h^k\delta\mathbf{u}_h^k\|_{\mathbf{L}^2}^2 + \|\sigma_h^{k+1}\delta^2\mathbf{u}_h^{k+1}\|_{\mathbf{L}^2}^2. \quad (4.72)$$

This estimate will be used repeatedly.

Now we proceed in two steps, as in the proof of Theorem 9: First we investigate the time steps $k = 1$, 2, then we investigate the cases $k \geq 3$.

(i) *Initialization:* Let $k \in \{1, 2\}$ and set $\mathbf{v}_h := 4\tau\mathbf{u}_h^{k+1}$ in (4.67). Using (4.72) and the Cauchy-Schwarz inequality we get,

$$(3-4M\tau)\|\sigma_h^{k+1}\mathbf{u}_h^{k+1}\|_{\mathbf{L}^2}^2 + 4\mu\tau\|\mathbf{u}_h^{k+1}\|_{\mathbf{H}^1}^2 \leq \frac{8\tau^2}{\chi}\|\nabla p_h^\sharp\|_{\mathbf{L}^2}^2 + \frac{\chi}{2}\|\sigma_h^{k+1}\mathbf{u}_h^{k+1}\|_{\mathbf{L}^2}^2,$$

which by (4.20) implies that if τ small enough

$$\|\sigma_h^{k+1}\mathbf{u}_h^{k+1}\|_{\mathbf{L}^2}^2 + 4\mu\tau\|\mathbf{u}_h^{k+1}\|_{\mathbf{H}^1}^2 \leq c\left(\|\sigma_h^0\mathbf{u}_h^0\|_{\mathbf{L}^2}^2 + \|\sigma_h^1\mathbf{u}_h^1\|_{\mathbf{L}^2}^2 + \frac{\tau^2}{\chi}\|\nabla p_h^0\|_{\mathbf{L}^2}^2 + \frac{\tau^2}{\chi}\|\nabla p_h^1\|_{\mathbf{L}^2}^2\right).$$

The estimate on the pressure is obtained *mutatis mutandis* the argument in the initialization step of the proof of Theorem 9. Hence

$$\|\sigma_h^{k+1}\mathbf{u}_h^{k+1}\|_{\mathbf{L}^2}^2 + 4\mu\tau\|\mathbf{u}_h^{k+1}\|_{\mathbf{H}^1}^2 + \frac{\tau^2}{\chi}\|\nabla p_h^{k+1}\|_{\mathbf{L}^2}^2 + \frac{\tau^2}{\chi}\|\nabla\delta p_h^{k+1}\|_{\mathbf{L}^2}^2 \leq c\left(\|\sigma_h^0\mathbf{u}_h^0\|_{\mathbf{L}^2}^2 + \|\sigma_h^1\mathbf{u}_h^1\|_{\mathbf{L}^2}^2 + \frac{\tau^2}{\chi}\|\nabla p_h^0\|_{\mathbf{L}^2}^2 + \frac{\tau^2}{\chi}\|\nabla p_h^1\|_{\mathbf{L}^2}^2\right), \quad k = 1, 2.$$

(ii) *General Step:* For $k \geq 3$ we proceed as in the general step for the constant density

case. Using (4.72) we obtain the estimate

$$
\begin{aligned}
(3-4M\tau)\|\sigma_h^{k+1}\mathbf{u}_h^{k+1}\|_{\mathbf{L}^2}^2-(4+8M\tau)\|\sigma_h^{k}\mathbf{u}_h^{k}\|_{\mathbf{L}^2}^2+(1-6M\tau)\|\sigma_h^{k-1}\mathbf{u}_h^{k-1}\|_{\mathbf{L}^2}^2\\
+2\|\sigma_h^{k+1}\delta\mathbf{u}_h^{k+1}\|_{\mathbf{L}^2}^2-2\|\sigma_h^{k}\delta\mathbf{u}_h^{k}\|_{\mathbf{L}^2}^2+\|\sigma_h^{k+1}\delta^2\mathbf{u}_h^{k+1}\|_{\mathbf{L}^2}^2+\\
4\mu\tau\|\mathbf{u}_h^{k+1}\|_{\mathbf{H}^1}^2+\frac{4\tau^2}{3\chi}\left[\|\nabla p_h^{k+1}\|_{\mathbf{L}^2}^2-\|\nabla p_h^{k}\|_{\mathbf{L}^2}^2+\|\nabla\delta p_h^{k}\|^2\right]\\
-\frac{4\tau^2}{3\chi}\|\nabla\delta^2 p_h^{k+1}\|_{\mathbf{L}^2}^2+\frac{8\tau^2}{9\chi}\langle\nabla\delta^2 p_h^{k},\nabla\delta p_h^{k+1}\rangle\leq 0.
\end{aligned}
$$

Add and subtract to this inequality the terms $2\chi\|\delta\mathbf{u}_h\|_{\mathbf{L}^2}^2$ taken at time steps t_{k+1} and t_k. Now, as in the constant density case, use the identity

$$
\chi\left\|\delta\mathbf{u}_h\right\|_{\mathbf{L}^2}^2=\left\|\chi^{1/2}\delta\mathbf{u}_h-\frac{2\tau}{3\chi^{1/2}}\nabla\delta^2 p_h\right\|_{\mathbf{L}^2}^2+\frac{4\tau^2}{9\chi}\left\|\nabla\delta^2 p_h\right\|_{\mathbf{L}^2}^2,
$$

to deduce

$$
\begin{aligned}
(3-4M\tau)\|\sigma_h^{k+1}\mathbf{u}_h^{k+1}\|_{\mathbf{L}^2}^2-(4+8M\tau)\|\sigma_h^{k}\mathbf{u}_h^{k}\|_{\mathbf{L}^2}^2+(1-6M\tau)\|\sigma_h^{k-1}\mathbf{u}_h^{k-1}\|_{\mathbf{L}^2}^2\\
+\|\sigma_h^{k+1}\delta^2\mathbf{u}_h^{k+1}\|_{\mathbf{L}^2}^2+4\mu\tau\|\mathbf{u}_h^{k+1}\|_{\mathbf{H}^1}^2\\
+2\left\|(\rho_h^{k+1}-\chi)^{1/2}\delta\mathbf{u}_h^{k+1}\right\|_{\mathbf{L}^2}^2-2\left\|(\rho_h^{k}-\chi)^{1/2}\delta\mathbf{u}_h^{k}\right\|_{\mathbf{L}^2}^2\\
+2\left\|\chi^{1/2}\delta\mathbf{u}_h^{k+1}-\frac{2\tau}{3\chi^{1/2}}\nabla\delta^2 p_h^{k+1}\right\|_{\mathbf{L}^2}^2-2\left\|\chi^{1/2}\delta\mathbf{u}_h^{k}-\frac{2\tau}{3\chi^{1/2}}\nabla\delta^2 p_h^{k}\right\|_{\mathbf{L}^2}^2\\
+\frac{4\tau^2}{3\chi}\left[\|\nabla p_h^{k+1}\|_{\mathbf{L}^2}^2-\|\nabla p_h^{k}\|_{\mathbf{L}^2}^2+\|\nabla\delta p_h^{k}\|_{\mathbf{L}^2}^2\right]\\
-\frac{4\tau^2}{9\chi}\|\nabla\delta^2 p_h^{k+1}\|_{\mathbf{L}^2}^2-\frac{8\tau^2}{9\chi}\|\nabla\delta^2 p_h^{k}\|_{\mathbf{L}^2}^2+\frac{8\tau^2}{9\chi}\langle\nabla\delta^2 p_h^{k},\nabla\delta p_h^{k+1}\rangle\leq 0, \quad (4.73)
\end{aligned}
$$

where we used assumption (4.20).

By assumption (4.20), the control on the last three pressure terms is obtained in

a similar way as in the proof of Theorem 9, thus giving

$$-\frac{4\tau^2}{9\chi}\|\nabla\delta^2 p_h^{k+1}\|_{\mathbf{L}^2}^2-\frac{8\tau^2}{9\chi}\|\nabla\delta^2 p_h^{k}\|_{\mathbf{L}^2}^2+\frac{8\tau^2}{9\chi}\langle\nabla\delta^2 p_h^{k},\nabla\delta p_h^{k+1}\rangle\geq$$
$$-\|\sigma_h^{k+1}\delta^2\mathbf{u}_h^{k+1}\|_{\mathbf{L}^2}^2+\frac{4\tau^2}{9\chi}\left[\|\nabla\delta p_h^{k}\|_{\mathbf{L}^2}^2-\|\nabla\delta p_h^{k-1}\|_{\mathbf{L}^2}^2\right].$$

Applying this estimate to (4.73) we arrive at the energy estimate

$$\begin{aligned}(3-4M\tau)\|\sigma_h^{k+1}\mathbf{u}_h^{k+1}\|_{\mathbf{L}^2}^2-(4+8M\tau)\|\sigma_h^{k}\mathbf{u}_h^{k}\|_{\mathbf{L}^2}^2+(1-6M\tau)\|\sigma_h^{k-1}\mathbf{u}_h^{k-1}\|_{\mathbf{L}^2}^2&\\
+4\mu\tau\|\mathbf{u}_h^{k+1}\|_{\mathbf{H}^1}^2&\\
+2\left\|(\rho_h^{k+1}-\chi)^{1/2}\delta\mathbf{u}_h^{k+1}\right\|_{\mathbf{L}^2}^2-2\left\|(\rho_h^{k}-\chi)^{1/2}\delta\mathbf{u}_h^{k}\right\|_{\mathbf{L}^2}^2&\\
+2\left\|\chi^{1/2}\delta\mathbf{u}_h^{k+1}-\frac{2\tau}{3\chi^{1/2}}\nabla\delta^2 p_h^{k+1}\right\|_{\mathbf{L}^2}^2-2\left\|\chi^{1/2}\delta\mathbf{u}_h^{k}-\frac{2\tau}{3\chi^{1/2}}\nabla\delta^2 p_h^{k}\right\|_{\mathbf{L}^2}^2&\\
+\frac{4\tau^2}{3\chi}\left[\|\nabla p_h^{k+1}\|_{\mathbf{L}^2}^2-\|\nabla p_h^{k}\|_{\mathbf{L}^2}^2+\|\nabla\delta p_h^{k}\|_{\mathbf{L}^2}^2\right]&\\
+\frac{4\tau^2}{9\chi}\left[\|\nabla\delta p_h^{k}\|_{\mathbf{L}^2}^2-\|\nabla\delta p_h^{k-1}\|_{\mathbf{L}^2}^2\right]\leq 0.&\quad(4.74)\end{aligned}$$

Introducing the notation

$$\begin{aligned}&A:=3-4M\tau,\quad B=-(4+8M\tau),\quad C=1-6M\tau,\\
&a^k:=\|\sigma_h^{k}\mathbf{u}_h^{k}\|_{\mathbf{L}^2}^2,\quad k\geq 0,\\
&b^k:=4\mu\tau\|\mathbf{u}_h^{k}\|_{\mathbf{L}^2}^2+\frac{4\tau^2}{3\chi}\|\nabla\delta p_h^{k-1}\|_{\mathbf{L}^2}^2,\quad k\geq 1,\\
&d^k:=2\left\|\left(\rho_h^{k}-\chi\right)^{1/2}\delta\mathbf{u}_h^{k}\right\|_{\mathbf{L}^2}^2+2\left\|\chi^{1/2}\delta\mathbf{u}_h^{k}+\frac{2\tau}{3\chi^{1/2}}\nabla\delta^2 p_h^{k}\right\|_{\mathbf{L}^2}^2\\
&\quad+\frac{4\tau^2}{3\chi}\|\nabla p_h^{k}\|_{\mathbf{L}^2}^2+\frac{4\tau^2}{9\chi}\|\nabla\delta p_h^{k-1}\|_{\mathbf{L}^2}^2,\quad k\geq 2,\end{aligned}$$

inequality (4.74) can be rewritten as

$$Aa^{k+1}+Ba^{k}+Ca^{k-1}\leq-\left(b^{k+1}+d^{k+1}-d^{k}\right),\quad k\geq 3.$$

Define $g^{k+1}:=-(b^{k+1}+d^{k+1}-d^k)$. If τ is small enough, this three-term recursion

inequality satisfies the assumptions of Proposition 12 of Appendix A. The roots of the characteristic polynomial are

$$r_1 := \frac{2+4M\tau-\sqrt{1+38M\tau-8M\tau^2}}{3-4M\tau} = \frac{1}{3}\left(1-\frac{41M\tau}{3}+\mathcal{O}(\tau^2)\right),$$
$$r_2 := \frac{2+4M\tau+\sqrt{1+38M\tau-8M\tau^2}}{3-4M\tau} = 1+9M\tau+\mathcal{O}(\tau^2).$$

Both roots are positive, the first one is strictly less than one third, and the second is greater but close to one. Hence, for $\nu \geq 3$

$$a^\nu \leq c(a^1+a^2)(r_1^\nu+r_2^\nu) - \frac{1}{3-4M\tau}\sum_{l=3}^{\nu} r_1^{\nu-l}\sum_{s=3}^{l} r_2^{l-s}(b^s+d^s-d^{s-1}),$$

which, since τ is small, can be rewritten as

$$a^\nu + \frac{1}{3}b^\nu \leq \mathcal{K}(1+e^{cT})(a^1+a^2) - \frac{1}{3-4M\tau}\sum_{l=3}^{\nu} r_1^{\nu-l}\sum_{s=3}^{l} r_2^{l-s}(d^s-d^{s-1}), \qquad (4.75)$$

for some constants c and $\mathcal{K}$.

Notice that

$$\sum_{s=3}^{l} r_2^{l-s}(d^s-d^{s-1}) = d^l + (r_2-1)\sum_{s=3}^{l-1} r_2^{l-s-1}d^s.$$

Hence (4.75) implies

$$a^\nu + \frac{1}{3}b^\nu + \frac{1}{3}d^\nu \leq \mathcal{K}(1+e^{cT})(a^1+a^2).$$

This inequality combined with the estimates obtained at the initialization step imply the result. □

Conjecture 2. As numerical experiments show (see Section F) the algorithm (4.67)-(4.69)-(4.70) performs as well as its constant density counterpart. This leads us to

believe that the following error estimates hold:

$$\|(\sigma \mathsf{u})_\tau - (\sigma_h \mathbf{u}_h)_\tau\|_{\ell^\infty(\mathbf{L}^2)} \leq c(\tau^2 + h^{l+1}),$$

and

$$\|\mathsf{u}_\tau - (\mathbf{u}_h)_\tau\|_{\ell^2(\mathbf{H}^1)} \leq c(\tau + h^l).$$

The techniques presented here, together with those of [39] may provide a proof of these facts.

Remark 37. In full analogy with the constant density case, it is possible to construct a rotational version (see [52, 82]) of the algorithm introduced above by replacing the pressure update (4.70) by the following: Find $p_h^{k+1} \in M_h$ so that,

$$\left\langle p_h^{k+1}, r_h \right\rangle = \left\langle p_h^k + \phi_h^{k+1}, r_h \right\rangle + \mu \left\langle \mathbf{u}_h^{k+1}, \nabla r_h \right\rangle. \tag{4.76}$$

The numerical experiments reported in Section F show that the algorithm (4.67)-(4.69)-(4.76) is stable and accurate.

F. Numerical Experiments

Convergence Tests

To test the accuracy of the second-order algorithm proposed in this paper, both in standard and rotational forms, we solve problem (1.3)-(1.4) using an analytical solution defined on the unit disk

$$\Omega = \{(x, y) \in \mathbb{R}^2 : x^2 + y^2 < 1\}. \tag{4.77}$$

The exact solution is

$$\rho(\mathbf{r},t) = 2 + r\cos(\theta - \sin t), \tag{4.78}$$

$$\mathsf{u}(\mathbf{r},t) = (-y,x)^\top \cos t, \tag{4.79}$$

$$\mathsf{p}(\mathbf{r},t) = \sin x \sin y \sin t, \tag{4.80}$$

and the corresponding right-hand side in the momentum equation is

$$\mathbf{f}(\mathbf{r},t) = \begin{pmatrix} (y\sin t - x\cos^2 t)\rho(\mathbf{r},t) + \cos x \sin y \sin t \\ -(x\sin t + y\cos^2 t)\rho(\mathbf{r},t) + \sin x \cos y \sin t \end{pmatrix}. \tag{4.81}$$

The computations are performed using the library `deal.II` (cf. [8, 7]). We use a $(\mathbb{Q}_2, \mathbb{Q}_2, \mathbb{Q}_1)$ approximation for the density, the velocity, and the pressure, respectively. We perform the accuracy tests with respect to τ on a mesh consisting of 5120 quadrangular cells. The dimensions of the vector spaces W_h, $\mathbf{X}_h$, and M_h are as follows:

$$\dim W_h = 20609, \tag{4.82}$$

$$\dim \mathbf{X}_h = 41218, \tag{4.83}$$

$$\dim M_h = 5185. \tag{4.84}$$

We measure the maximum over the time interval $[0, 10]$ of the errors measured in various norms. This mesh is chosen, so that the discretization error in space is significantly smaller than that induced by the time discretization. The convergence with respect to τ is verified in the range $5.10^{-3} \leq \tau \leq 1.10^{-1}$.

We test the second-order standard formulation described in Section E. The results are shown in Table VI. As expected, the error on the velocity and the density in the L^2-norm is of $\mathcal{O}(\tau^2)$ and the error on the velocity in the H^1-norm and on the

Table VI. Error in Time for Standard Scheme

τ	ρ–L^2	**Rate**	u–$\mathbf{L}^2$	**Rate**	u–$\mathbf{H}^1$	**Rate**	p–L^2	**Rate**
0.1	9.15E-003	—	6.93E-003	—	3.29E-002	—	4.34E-002	—
0.05	1.27E-003	2.84	1.70E-003	2.03	9.93E-003	1.73	1.21E-002	1.84
0.03	2.10E-004	2.60	4.20E-004	2.02	3.20E-003	1.64	3.62E-003	1.74
0.01	4.18E-005	2.33	1.05E-004	2.00	1.11E-003	1.52	1.19E-003	1.60
0.01	8.65E-006	2.27	2.61E-005	2.00	3.63E-004	1.62	3.78E-004	1.66

pressure in the L^2-norm is of $\mathcal{O}(\tau)$.

Next we test the rotational version of the method which consists of using the pressure update (4.76), introduced in Remark 37, instead of (4.70). The results are shown in Table VII. We observe that all the errors are fully second-order with respect to τ. It is likely that there is a super-convergence effect due to the regularity of the domain. We recall that a similar super-converge effect is observed for the rotational variant of the pressure-correction algorithm for constant density flows (see [52]). We conjecture that in general domains the error on the velocity measured in the L^2-norm is $\mathcal{O}(\tau^2)$, and the error on the velocity in the H^1-norm and on the pressure in the L^2-norm is $\mathcal{O}(\tau^{3/2})$.

The Rayleigh–Taylor Instability

We now illustrate the performance of the method on a realistic problem. We compute the development of a Rayleigh–Taylor instability in the viscous regime as documented by Tryggvason in [83]. This problem consists of two layers of fluid initially at rest

Table VII. Error in Time for Rotational Scheme

τ	ρ–L^2	**Rate**	u–$\mathbf{L}^2$	**Rate**	u–$\mathbf{H}^1$	**Rate**	p–L^2	**Rate**
0.1	3.70E-003	—	3.90E-003	—	1.59E-002	—	1.12E-002	—
0.05	6.38E-004	2.54	1.18E-003	1.73	4.89E-003	1.70	3.31E-003	1.76
0.03	1.35E-004	2.24	3.34E-004	1.82	1.43E-003	1.78	9.34E-004	1.83
0.01	3.21E-005	2.07	9.03E-005	1.89	4.03E-004	1.82	2.53E-004	1.88
0.01	7.85E-006	2.03	2.37E-005	1.93	1.12E-004	1.84	6.71E-005	1.92

in the rectangular domain $\Omega = (-d/2, d/2) \times (-2d, 2d)$. The transition between the two fluids is regularized as follows

$$\frac{\rho(x, y, t = 0)}{\rho_0^{\min}} = 2 + \tanh\left(\frac{y - \eta(x)}{0.01d}\right), \tag{4.85}$$

where the initial position of the perturbed interface is $\eta(x) = -0.1d\cos(2\pi x/d)$. The heavy fluid is above and the density ratio is 3, so that the Atwood number

$$A_t = \left(\rho_0^{\max} - \rho_0^{\min}\right) / \left(\rho_0^{\max} + \rho_0^{\min}\right), \tag{4.86}$$

equals 0.5, according to Tryggvason's definition, where we set $\rho_0^{\max} := \max_{\mathbf{x}\in\Omega} \rho_0(\mathbf{x})$. For $t > 0$ the system evolves under the action of a vertical downward gravity field of intensity $\mathbf{g}$; the source term in the momentum equation is downward and equal to ρg.

The equations are non-dimensionalized using the following references: $\rho_0^{\min}$ for the density, d for lengths, and $d^{1/2}/g^{1/2}$ for time, where g is the gravity field. Then, the reference velocity is $d^{1/2}g^{1/2}$, and the Reynolds number is defined by $Re = \rho_0^{\min} d^{3/2} g^{1/2}/\mu$. The computational domain can be restricted to $(0, d/2) \times (-2d, 2d)$

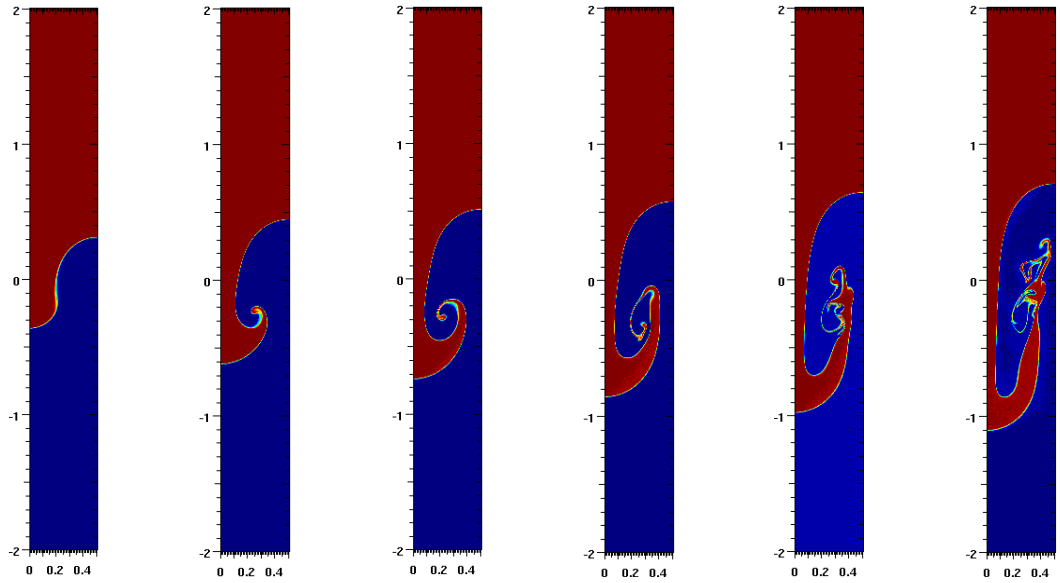

Fig. 2. Rayleigh-Taylor Instability. $Re = 1000$; density ratio 3. The interface is shown at times 1, 1.5, 1.75, 2, 2.25, and 2.5

since we assume that the symmetry of the initial condition is maintained during the time evolution. The no-slip condition is enforced at the bottom and top walls and symmetry is imposed on the two vertical sides.

The mass conservation equation is stabilized by adding a nonlinear viscosity proportional to the residual of the conservation equation for ρ^2 in the spirit of the entropy viscosity of [45].

The time evolution of the density field at $Re = 1000$ is shown in Fig. 2 at times 1, 1.5, 1.75, 2, 2.25, and 2.5 in the time scale of Tryggvason, which is related to ours by $t_{\text{Tryg}} = t\sqrt{A_t}$. The mesh is such that there are 466573 degrees of freedom for each component of the velocity. The mesh size is of order 0.025 in the refined regions. The time step is $\tau = 0.00125\sqrt{A_t}$.

To further assess the sensitivity of the method to spatial resolution and to verify that the numerical viscosity is significantly smaller than the physical viscosity we solve the same problem using the same mesh for $Re = 5000$. The results are shown

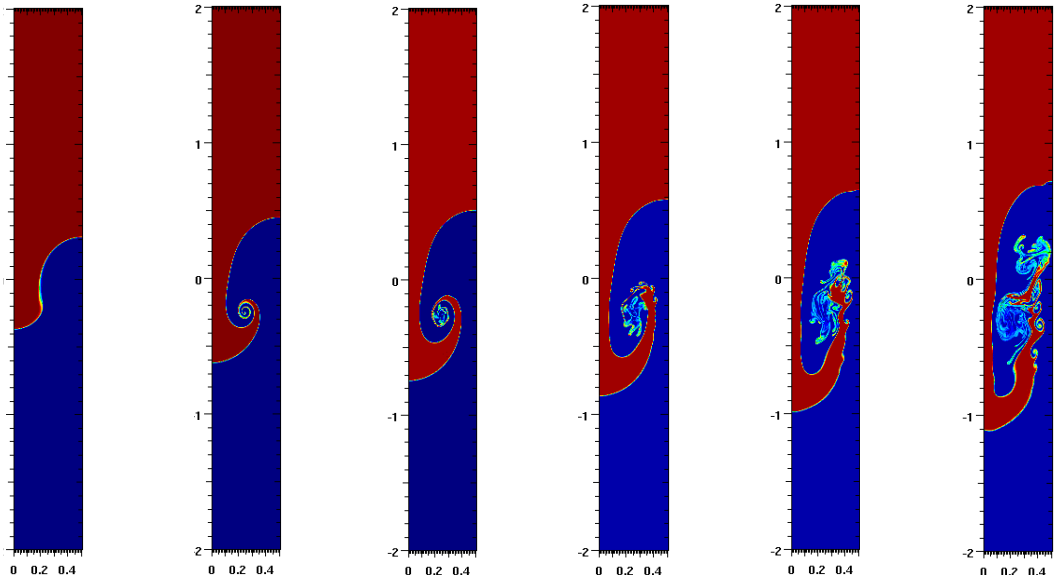

Fig. 3. Rayleigh-Taylor Instability. $Re = 5000$; density ratio 3. The interface is shown at times 1, 1.5, 1.75, 2, 2.25, and 2.5

in Fig. 3.

The above results are in good agreement with those from [31]. Since the algorithm of Section E only requires solving a Poisson equation, computing the above test cases was significantly faster (one order of magnitude) than when doing the computations reported in [31]. This time saving allowed us to use finer space resolution.

Next, we perform the test case reported in [11]. The geometry is the same as above. The density ratio is 7 so that $A_t = 0.75$, using Tryggvason's definition (4.86) (using the definition from [11] the Atwood number is 0.875). The initial density field is regularized as follows:

$$\frac{\rho(x, y, t = 0)}{\rho_0^{\min}} = 4 + 3 \tanh\left(\frac{y - \eta(x)}{0.01d}\right), \tag{4.87}$$

where the perturbation of the interface is given by $\eta(x) = -0.01d \cos(2\pi x/d)$. The Reynolds number is $Re = 1000$.

The results using the same mesh and same time step as in for the low density ratio

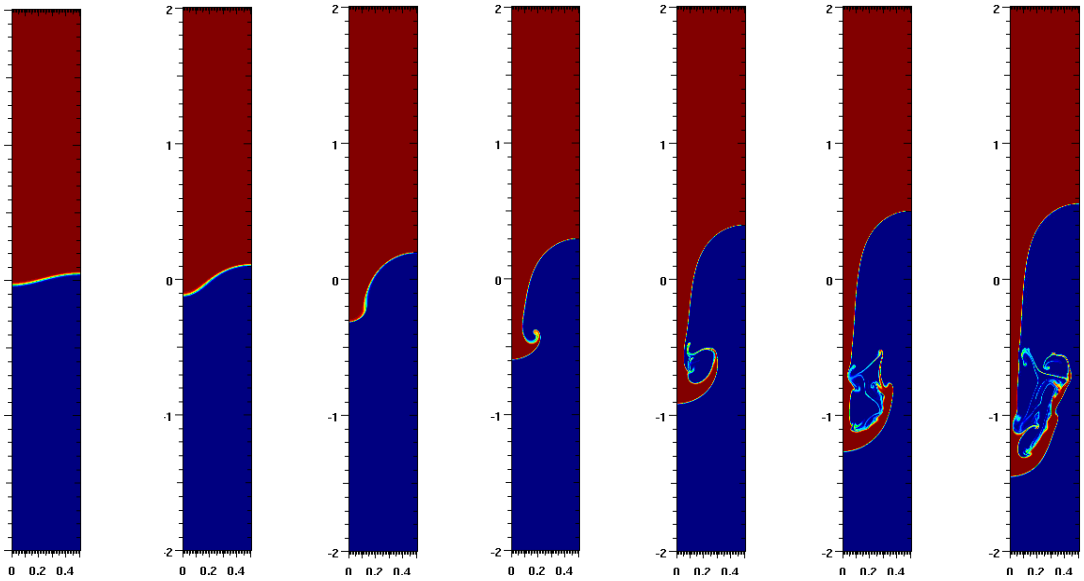

Fig. 4. Rayleigh-Taylor Instability. $Re = 1000$; density ratio 7. The interface is shown at times 1, 1.5, 2, 2.5, 3, 3.5, and 3.75

are reported in Fig. 4 for times 1, 1.5, 2, 2.5, 3, 3.5, and 3.75 (using $d^{1/2}/g^{1/2}$ as time scale). Although the locations of the falling and rising bubbles are similar to those reported in [11], the details of the flow differ from those in [11]. This unexplained discrepancy was already noted in [48].

A Lighter Bubble Rising in a Heavier Medium

Let us consider another realistic example. In a rectangular domain $\Omega = (-3d, 3d) \times (0, 9d)$ there is initially a bubble of fluid of radius d (with density ρ_1 and viscosity μ_1) immersed in a heavier medium (with density ρ_2 and viscosity μ_2). The system evolves under the action of a gravity field pointing downward and of intensity g. We non-dimensionalize the equations with the following references: ρ_1 for the density, $\sqrt{d/g}$ for time and d for lengths. The reference velocity is $\sqrt{dg}$ and the non-dimensional viscosities are computed as

$$\hat{\mu} = \frac{\mu}{\rho_1 d^{3/2} g^{1/2}}.$$

To properly model the relevant physics of the system it is necessary to include the surface tension effects. This, being a force that acts only on the interface between the two fluids, is quite complicated to properly handle numerically since it requires a good representation of the interface boundary. Several approaches have been proposed to handle such difficulty. Without being exhaustive, we can mention grid alignment techniques [9], moving mesh methods [75], level set methods [64, 65, 84], surface tracking [68] and phase field [10, 14, 56, 62] and [76]. Here we adopt the phase field approach of [76].

The idea of the phase field model is to replace the sharp interface between the fluids by a smooth transition layer of thickness η. Then it turns out that the evolution of the phase variable ϕ, which serves as a marker for each one of the phases, is given by the Cahn-Hillard equation

$$\phi_t + \mathsf{u}\cdot\nabla\phi = -\gamma\Delta(\Delta\phi - f(\phi)),$$

where $f = F'$ and F is the Ginzburg-Landau double well potential

$$F(\phi) = \frac{1}{4\eta^2}\left(\phi^2 - 1\right)^2.$$

However, the Cahn-Hillard equation involves fourth order derivatives, which are difficult to handle using finite elements. Therefore, the evolution law for the phase variable is usually replaced by the Allen-Cahn equation

$$\phi_t + \mathsf{u}\cdot\nabla\phi = \gamma(\Delta\phi - f(\phi)).$$

Then, the surface tension appears as a volume term in the momentum equation

$$\rho\mathsf{u}_t + \rho\mathsf{u}\cdot\nabla\mathsf{u} - \mu\Delta\mathsf{u} + \nabla\mathsf{p} + \lambda\nabla\cdot(\nabla\phi\otimes\nabla\phi) = \mathbf{f},$$

where λ is the mixing energy density. For details, the reader is referred to the sources

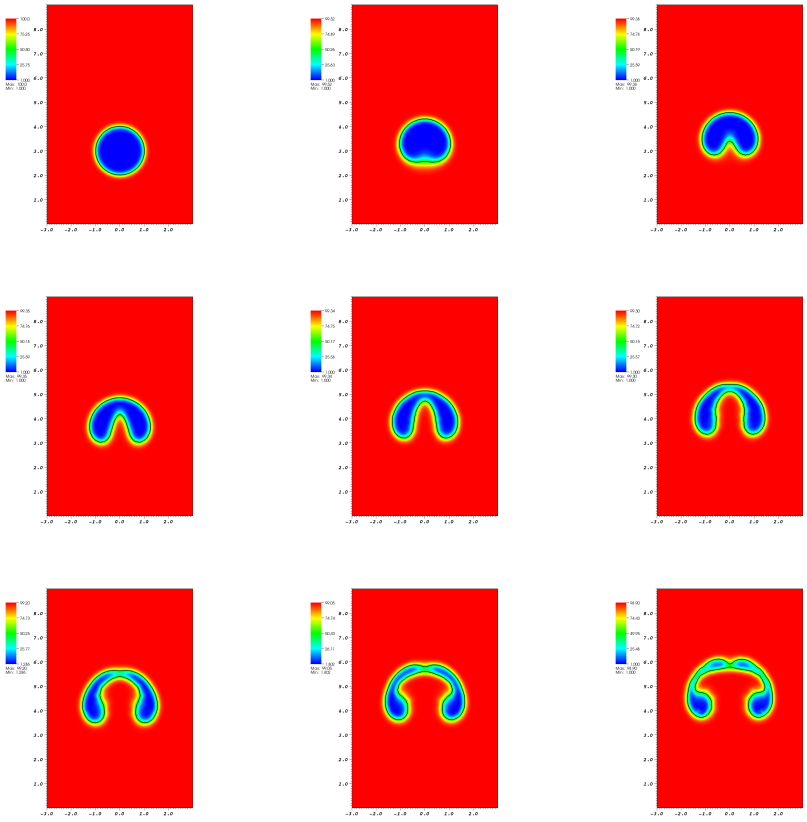

Fig. 5. Rising Bubble. $Re = 1000$; density ratio 100. The interface is shown at times 0, 1, 1.5, 2, 2.5, 3, 3.5, 4 and 4.5

cited above.

Let us consider the case where the density ratio $\rho_2/\rho_1 = 100$ and the two fluids have the same viscosity $\mu = 10^{-3}$. The space discretization of the problem is done using $(\mathbb{Q}_2, \mathbb{Q}_2, \mathbb{Q}_1)$ elements for the density, velocity and pressure, respectively. The mesh is uniform and it has 4480 rectangular cells, so that there are 18193 $\mathbb{Q}_2$-degrees of freedom and 4617 $\mathbb{Q}_1$-degrees of freedom. The time step is $\tau = 10^{-3}$. The interface thickness is taken equal to the mesh size. The results are shown in Figure 5, where we can see the interface at times 0, 1, 1.5, 2, 2.5, 3, 3.5, 4 and 4.5. The results are in good

agreement with similar ones obtained using different techniques (see the references cited above).

A Falling Drop

As a final example, let us consider a falling drop. The geometry is the same as before but, in this case, the subdomain $\Omega_{pool} = (-3d, 3d) \times (0, 3d)$ is filled with a heavy medium of density ρ_2. There is a circular drop of this same medium of radius d located at $(0, 6d)$. The rest of the domain is filled with a lighter medium of density ρ_1. The system is at rest initially and we follow its evolution under the action of gravity.

We non-dimensionalize the equations using the same references as in the previous example and consider the case $\rho_2/\rho_1 = 100$, with $\mu_2 = \mu_1 = 10^{-3}$. The mesh is as in the rising bubble and the time step is $\tau = 10^{-3}$.

To take into account the surface tension effects, we use the phase field method described above. The parameters are the same as for the rising bubble experiment.

A plot of the interface, together with the velocity field can be seen in Figure 6. Although the results are far from depicting all the details of the real phenomena (see [81], for instance), at least we are able to capture some of the most significant features of the phenomenon. It is possible that a combination of this method with more sophisticated schemes to take care of the interface will provide more accurate results. We leave this study for further investigation.

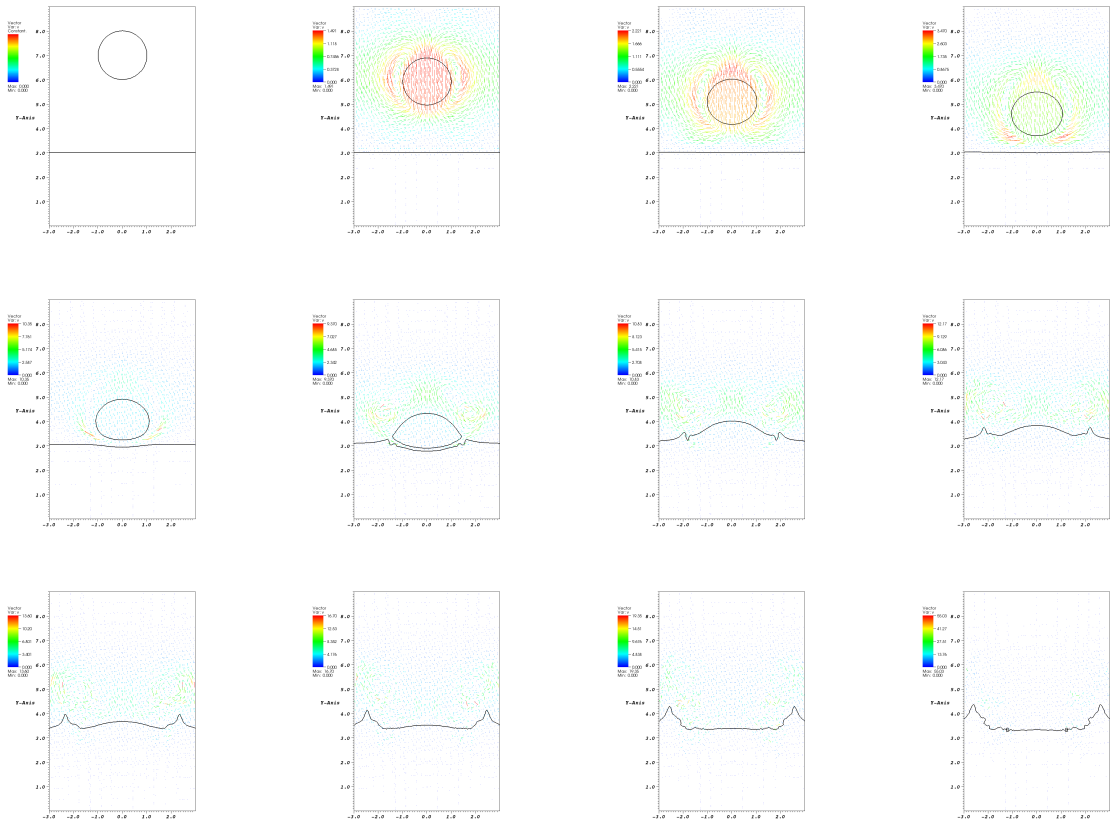

Fig. 6. Falling Drop. $Re = 1000$; density ratio 100. The interface is shown at times 0, 1.5, 2, 2.25, 2.5, 2.75, 2.9, 3, 3.1, 3.2, 3.3 and 3.35

CHAPTER V

CONCLUSION

During the course of this dissertation we have studied two models that arise in the study of complex fluid flow phenomena. For each one of them we have proposed effective discretization techniques and proved that they converge to the solution. Let us briefly review the obtained results.

For the nonlinear Darcy equations of Chapter III, in the case where the permeability is a bounded from above and a strictly positive function of the pressure, we have obtained sufficient conditions for a solution to be nonsingular. In the case of a unique solution, we have proposed a discretization scheme and we have proved optimal error estimates for this scheme. Moreover, we proposed an algorithm for the solution of this discrete system and we proved that this algorithm converges independently of the discretization parameter. In the case where there is no unique solution, we have proposed a discretization scheme for the approximation of nonsingular solutions. We have shown that this discretization scheme has optimal error estimates. Finally, we studied the convergence of a Newton type algorithm for the solution of the discrete system that approximates a nonsingular solution. We have shown that this method converges quadratically, but not uniformly with respect to the discretization parameter. This type of deterioration has been observed in several other problems.

In the case when the dependence of the drag coefficient on the pressure is exponential, we proposed a splitting scheme which requires the solution of two linear problems for the determination of the unknowns. Although the complete mathematical analysis the problem in this case remains an open question, under the assumption that there is a solution, we have showed that this splitting scheme converges to the solution. The obtained estimates are suboptimal, but the numerical experiments show

that this method is indeed optimal. A more refined analysis may provide a proof of this fact.

Concerning the approximation of incompressible viscous flows with variable density (see Chapter IV) we have proposed a new fractional time-stepping technique which decouples the diffusion and incompressibility constraint. The main novelty of this scheme lies in the fact that for the determination of the pressure one has to solve a Poisson equation, as opposed to a variable-coefficient second-order elliptic equation. This simplification greatly reduces the overall computational cost of the scheme, which allows for the use of finer meshes and smaller time steps.

We have proposed a family of first order schemes, and have shown that these schemes are stable, convergent and perform at least as good as their well-known counterparts used in the solution of constant density flows. Moreover, we have proposed a formally second order scheme and we proved its stability. Numerical experiments show that this scheme is indeed second order accurate. The techniques developed in this dissertation may enable us to prove this. However, we have not pursued this direction. Finally, as a byproduct of our analysis, we have provided a new proof of an old result. Namely, the stability of the so-called pressure correction incremental fractional time-stepping scheme in standard form. The novelty in our proof technique is that we have completely removed the solenoidal velocity from the analysis. This new family of methods has already proved useful in the development of new and simpler fractional time-stepping schemes for incompressible flows. For instance, [76] uses these ideas to introduce numerical methods for a phase-field model for two-phase flows. Moreover, the ideas and techniques that we have here introduced, have served as a basis for the development of a new class of methods for the Navier-Stokes equations based on direction splitting. The reader is referred to [42, 41, 43] for details.

REFERENCES

[1] Y. Achdou and C. Bernardi, *Adaptive finite volume or finite element discretization of Darcy's equations with variable permeability*, C. R. Acad. Sci. Paris, Sér. I, 333 (2001), pp. 693–698.

[2] R. A. Adams, *Sobolev spaces*, Pure and Appl. Math., 65, Academic Press, New York-London, 1975.

[3] G. Allaire, *Homogenization of the Navier-Stokes equations with a slip boundary condition*, Comput. Methods Appl. Mech. Engrg., 44 (1991), pp. 605–641.

[4] A. S. Almgren, J. B. Bell, P. Colella, L. H. Howell, and M. L. Welcome, *A conservative adaptive projection method for the variable density incompressible Navier-Stokes equations*, J. Comput. Phys., 142 (1998), pp. 1–46.

[5] M. Azaïez, F. Ben Belgacem, C. Bernardi, and N. Chorfi, *Spectral discretization of Darcy's equations with pressure dependent porosity*, Tech. Rep. R09010, Laboratoire Jacques-Louis Lions, 2009. `http://www.ann.jussieu.fr/publications/2009/R09010.html`.

[6] I. Babuška, *The finite element method with Lagrangian multipliers*, Numer. Math., 20 (1973), pp. 179–192.

[7] W. Bangerth, R. Hartmann, and G. Kanschat, *deal.II Differential Equations Analysis Library, Technical Reference*. `http://www.dealii.org` accessed on 5/26/2010.

[8] W. Bangerth, R. Hartmann, and G. Kanschat, *deal.II—a general-purpose object-oriented finite element library*, ACM Trans. Math. Software, 33 (2007), pp. 24–27.

[9] B. Bejanov, J. L. Guermond, and P. D. Minev, *A grid-alignment finite element technique for incompressible multicomponent flows*, J. Comput. Phys., 227 (2008), pp. 6473–6489.

[10] A. Béliveau, A. Fortin, and Y. Demay, *A two-dimensional numerical method for the deformation of drops with surface tension*, Int. J. Comput. Fluid Dyn., 10 (1998), pp. 225–240.

[11] J. B. Bell and D. L. Marcus, *A second-order projection method for variable-density flows*, J. Comput. Phys., 101 (1992), pp. 334–348.

[12] J. Bergh and J. Löfström, *Interpolation Spaces. An Introduction*, Grundlehren der Mathematischen Wissenschaften, 223, Springer-Verlag, Berlin, 1976.

[13] D. Boffi, F. Brezzi, L. F. Demkowicz, R. G. Durán, R. S. Falk, and M. Fortin, *Mixed finite elements, compatibility conditions, and applications*, Lect. Notes Math., 1939, Springer-Verlag, Berlin, 2008.

[14] J. U. Brackbill, D. B. Kothe, and C. Zemach, *A continuum method for modeling surface tension*, J. Comput. Phys., 100 (1992), pp. 335–354.

[15] S. C. Brenner and L. R. Scott, *The Mathematical Theory of Finite Element Methods*, Texts in Appl. Math. 15, Springer, New York, 1994.

[16] F. Brezzi and M. Fortin, *Mixed and Hybrid Finite Element Methods*, Springer Series Comput. Math., 15, Springer-Verlag, New York, 1991.

[17] F. Brezzi, J. Rappaz, and P.-A. Raviart, *Finite-dimensional approximation of nonlinear problems. I. Branches of nonsingular solutions*, Numer. Math., 36 (1980/81), pp. 1–25.

[18] E. Burman and P. Hansbo, *Edge stabilization for Galerkin approximations of convection-diffusion-reaction problems*, Comput. Methods Appl. Mech. Engrg., 193 (2004), pp. 1437–1453.

[19] L. Cattabriga, *Su un problema al contorno relativo al sistema di equazioni di Stokes*, Rend. Sem. Mat. Univ. Padova, 31 (1961), pp. 308–340.

[20] A. J. Chorin, *Numerical solution of the Navier-Stokes equations*, Math. Comp., 22 (1968), pp. 745–762.

[21] P. G. Ciarlet, *Basic error estimates for elliptic problems*, in Handbook of numerical analysis, Vol. II, North-Holland, Amsterdam, 1991, pp. 17–351.

[22] D. Cioranescu, P. Donato, and H. I. Ene, *Homogenization of the Stokes problem with non-homogeneous slip boundary conditions*, Math. Methods Appl. Sci., 19 (1996), pp. 857–881.

[23] H. Darcy, *Les Fontaines Publiques de la Ville de Dijon*, Victor Dalmont, Paris, 1856.

[24] J. Douglas, Jr. and T. Dupont, *A Galerkin method for a nonlinear Dirichlet problem*, Math. Comp., 29 (1975), pp. 689–696.

[25] J. Douglas, Jr. and T. F. Russell, *Numerical methods for convection-dominated diffusion problems based on combining the method of characteristics with finite element or finite difference procedures*, SIAM J. Numer. Anal., 19 (1982), pp. 871–885.

[26] H. I. Ene and E. Sánchez-Palencia, *Équations et phénomènes de surface pour l'écoulement dans un modèle de milieu poreux*, J. Mécanique, 14 (1975), pp. 73–108.

[27] A. Ern and J.-L. Guermond, *Theory and Practice of Finite Elements*, Appl. Math. Sci., 159, Springer-Verlag, New York, 2004.

[28] E. Fernández Cara and F. Guillén, *Some new existence results for the variable density Navier-Stokes equations*, Ann. Fac. Sci. Toulouse Math. (6), 2 (1993), pp. 185–204.

[29] G. B. Folland, *Real analysis*, Pure and Appl. Math., John Wiley & Sons Inc., New York, second ed., 1999.

[30] P. Forchheimer, *Wasserbewegung durch Boden*, Z. Ver. Deutsh. Ing., 45 (1901), pp. 1782–1788.

[31] Y. Fraigneau, J.-L. Guermond, and L. Quartapelle, *Approximation of variable density incompressible flows by means of finite elements and finite volumes*, Comput. Methods Appl. Mech. Engrg., 17 (2001), pp. 893–902.

[32] V. Girault, F. Murat, and A. Salgado, *Finite element discretization of Darcy's equations with pressure dependent porosity*, M2AN Math. Model. Numer. Anal., in press (2010). DOI: 10.1051/m2an/2010019.

[33] V. Girault, R. H. Nochetto, and L. R. Scott, *Maximum-norm stability of the finite element Stokes projection*, J. Math. Pures Appl., 84 (2005), pp. 279–330.

[34] V. Girault and P.-A. Raviart, *Finite Element Methods for Navier-Stokes Equations. Theory and Algorithms*, Springer Series in Comput. Math., 5, Springer-Verlag, Berlin, 1986.

[35] V. Girault and M. F. Wheeler, *Numerical discretization of a Darcy-Forchheimer model*, Numer. Math., 110 (2008), pp. 161–198.

[36] P. Grisvard, *Elliptic problems in Nonsmooth Domains*, Monographs and Studies in Mathematics, 24, Pitman, Boston, 1985.

[37] J.-L. Guermond, *Sur l'approximation des équations de Navier–Stokes par une méthode de projection*, C. R. Acad. Sci. Paris, Sér. I, 319 (1994), pp. 887–892.

[38] J.-L. Guermond, *Some practical implementations of projection methods for Navier–Stokes equations*, M2AN Math. Model. Numer. Anal., 30 (1996), pp. 637–667.

[39] J.-L. Guermond, *Un résultat de convergence d'ordre deux en temps pour l'approximation des équations de Navier-Stokes par une technique de projection incrémentale*, M2AN Math. Model. Numer. Anal., 33 (1999), pp. 169–189.

[40] J.-L. Guermond, A. Marra, and L. Quartapelle, *Subgrid stabilized projection method for 2D unsteady flows at high Reynolds numbers*, Comput. Methods Appl. Mech. Engrg., 195 (2006), pp. 5857–5876.

[41] J.-L. Guermond and P. D. Minev, *A new class of fractional step techniques for the incompressible Navier-Stokes equations using direction splitting*, C. R. Acad. Sci. Paris, Sér. I, 348 (2010), pp. 581–585.

[42] J.-L. Guermond and P. D. Minev, *A new class of splitting methods for the incompressible Navier-Stokes equations using direction splitting.* submitted to J. Comput. Phys., 2010.

[43] J.-L. Guermond, P. D. Minev, and A. J. Salgado, *Convergence analysis of a new class of direction splitting algorithms for the Navier-Stokes equations.* In preparation, 2010.

[44] J.-L. Guermond, P. D. Minev, and J. Shen, *An overview of projection methods for incompressible flows*, Comput. Methods Appl. Mech. Engrg., 195 (2006), pp. 6011–6045.

[45] J.-L. Guermond and R. Pasquetti, *Entropy-based nonlinear viscosity for Fourier approximations of conservation laws*, C. R. Math. Acad. Sci. Paris, 346 (2008), pp. 801–806.

[46] J.-L. Guermond and L. Quartapelle, *Calculation of incompressible viscous flows by an unconditionally stable projection FEM*, J. Comput. Phys., 132 (1997), pp. 12–33.

[47] J.-L. Guermond and L. Quartapelle, *On the approximation of the unsteady Navier–Stokes equations by finite element projection methods*, Numer. Math., 80 (1998), pp. 207–238.

[48] J.-L. Guermond and L. Quartapelle, *A projection FEM for variable density incompressible flows*, J. Comput. Phys., 165 (2000), pp. 167–188.

[49] J.-L. Guermond and A. Salgado, *A fractional step method based on a pressure poisson equation for incompressible flows with variable density*, C. R. Acad. Sci. Paris, Sér. I, 346 (2008), pp. 913 – 918.

[50] J.-L. Guermond and A. Salgado, *Error analysis of a fractional time-stepping technique for incompressible flows with variable density.* submitted to SIAM J. Numer. Anal., 2009.

[51] J.-L. Guermond and A. Salgado, *A splitting method for incompressible flows with variable density based on a pressure poisson equation*, J. Comput. Phys., 228 (2009), pp. 2834 – 2846.

[52] J. L. Guermond and J. Shen, *On the error estimates for the rotational pressure-correction projection methods*, Math. Comp., 73 (2004), pp. 1719–1737.

[53] F. Hecht, A. Le Hyaric, O. Pironneau, and K. Ohtsuka, *Freefem++*, Laboratoire Jacques-Louis Lions, Université Pierre et Marie Curie, Paris, 2008.

[54] A. Y. Helemskii, *Lectures and exercises on functional analysis*, Transl. of Math. Mon., 233, American Mathematical Society, Providence, 2006. Translated from the 2004 Russian original by S. Akbarov.

[55] J. Heywood and R. Rannacher, *Finite element approximation of the nonstationary Navier-Stokes problem. I. Regularity of solutions and second-order error estimates for spatial discretization*, SIAM J. Numer. Anal., 19 (1982), pp. 275–311.

[56] D. Jacqmin, *Calculation of two-phase Navier-Stokes flows using phase-field modeling*, J. Comput. Phys., 155 (1999), pp. 96–127.

[57] C. Johnson, U. Nävert, and J. Pitkäranta, *Finite element methods for linear hyperbolic equations*, Comput. Methods Appl. Mech. Engrg., 45 (1984), pp. 285–312.

[58] L. V. Kantorovich and G. P. Akilov, *Funktsionalnyi Analiz*, Nauka, Moscow, third ed., 1984.

[59] D. Kim and E.-J. Park, *Primal mixed finite-element approximation of elliptic equations with gradient nonlinearities*, Comput. Math. Appl., 51 (2006), pp. 793–804.

[60] J.-L. Lions and E. Magenes, *Problèmes aux Limites non Homogènes et Applications*, vol. 1, Dunod, Paris, France, 1968.

[61] P.-L. Lions, *Mathematical Topics in Fluid Mechanics. vol. 1. Incompressible Models*, Oxford Lect. Ser. in Math. and its Appl., 3, Oxford University Press, New York, 1996.

[62] C. Liu and J. Shen, *A phase field model for the mixture of two incompressible fluids and its approximation by a Fourier-spectral method*, Phys. D, 179 (2003), pp. 211–228.

[63] C. Liu and N. Walkington, *Convergence of numerical approximations of the incompressible Navier-Stokes equations with variable density and viscosity*, SIAM J. Numer. Anal., 45 (2007), pp. 1287–1304.

[64] E. Olsson and G. Kreiss, *A conservative level set method for two phase flow*, J. Comput. Phys., 210 (2005), pp. 225–246.

[65] E. Olsson, G. Kreiss, and S. Zahedi, *A conservative level set method for two phase flow. II*, J. Comput. Phys., 225 (2007), pp. 785–807.

[66] E.-J. Park, *Mixed finite element methods for nonlinear second-order elliptic problems*, SIAM J. Numer. Anal., 32 (1995), pp. 865–885.

[67] S. E. Pastukhova, *Justification of Darcy's law for a porous medium with an incomplete no-slip condition*, Mat. Sb., 189 (1998), pp. 135–153.

[68] N. Périnet, D. Juric, and L. Tuckerman, *Numerical simulation of Faraday waves*, J. Fluid Mech., 635 (2009), pp. 1–26.

[69] J.-H. Pyo and J. Shen, *Gauge-Uzawa methods for incompressible flows with variable density*, J. Comput. Phys., 221 (2007), pp. 181–197.

[70] K. R. Rajagopal, *On a hierarchy of approximate models for flows of incompressible fluids through porous solids*, Math. Models Methods Appl. Sci., 17 (2007), pp. 215–252.

[71] R. Rannacher, *On Chorin's projection method for the incompressible Navier-Stokes equations*, in The Navier-Stokes equations II—Theory and Numerical Methods (Oberwolfach, Germany 1991), Lecture Notes in Math., 1530, Springer, Berlin, 1992, pp. 167–183.

[72] J. E. Roberts and J.-M. Thomas, *Mixed and hybrid methods*, in Handbook of numerical analysis, Vol. II, North-Holland, Amsterdam, 1991, pp. 523–639.

[73] J. Schöberl and W. Zulehner, *Symmetric indefinite preconditioners for saddle point problems with applications to PDE-constrained optimization problems*, SIAM J. Matrix Anal. Appl., 29 (2007), pp. 752–773.

[74] J. Shen, *On error estimates of projection methods for the Navier-Stokes equations: First-order schemes*, SIAM J. Numer. Anal., 29 (1992), pp. 57–77.

[75] J. Shen and X. Yang, *An efficient moving mesh spectral method for the phase-field model of two-phase flows*, J. Comput. Phys., 228 (2009), pp. 2978–2992.

[76] J. Shen and X. Yang, *A phase-field model and its numerical approximation for two-phase incompressible flows with different densities and viscosities*, SIAM J. Sci. Comput, 32 (2010), pp. 1159–1179.

[77] E. Skjetne and J.-L. Auriault, *Homogenization of wall-slip gas flow through porous media*, Transp. Porous Media, 36 (1999), pp. 293–306.

[78] L. Tartar, *An Introduction to Sobolev Spaces and Interpolation Spaces*, Lect. Notes of the Unione Matematica Italiana, 3, Springer, Berlin, Germany, 2007.

[79] R. Temam, *Sur l'approximation de la solution des équations de Navier-Stokes par la méthode des pas fractionnaires II*, Arch. Rat. Mech. Anal., 33 (1969), pp. 377–385.

[80] R. Temam, *Navier-Stokes Equations. Theory and Numerical Analysis*, AMS Chelsea Publishing, Providence, RI, 2001. Reprint of the 1984 edition.

[81] S. T. Thoroddsen, T. G. Etoh, and K. Takehara, *High-speed imaging of drops and bubbles*, in Annu. Rev. Fluid Mech., 40, Annual Reviews, Palo Alto, CA, 2008, pp. 257–285.

[82] L. J. P. Timmermans, P. D. Minev, and F. N. van de Vosse, *An approximate projection scheme for incompressible flow using spectral elements*, Int. J. Numer. Methods Fluids, 22 (1996), pp. 673–688.

[83] G. Tryggvason, *Numerical simulation of Rayleigh–Taylor instability*, J. Comput. Phys., 75 (1988), pp. 253–282.

[84] L. Ville, L. Silva, and T. Coupez, *Convected level set method for the numerical simulation of fluid buckling*, Int. J. Numer. Methods Fluids, in press (2010). DOI 10.1002/fld.2259.

[85] N. J. Walkington, *Convergence of the discontinuous Galerkin method for discontinuous solutions*, SIAM J. Numer. Anal., 42 (2004), pp. 1801–1817.

[86] M. Wheeler, *A priori L_2 error estimates for Galerkin approximations to parabolic partial differential equations*, SIAM J. Numer. Anal., 10 (1973), pp. 723–759.

[87] W. Zulehner, *Analysis of iterative methods for saddle point problems: A unified approach*, Math. Comp., 71 (2002), pp. 479–505.

APPENDIX A

THREE TERM RECURSION INEQUALITIES

Let us prove auxiliary results regarding three term recursion inequalities. These results will be needed to prove stability of the algorithms (4.14)–(4.15)–(4.13) and (4.65)–(4.70).

Proposition 11. *Assume that the characteristic polynomial of the three term recursion equation*

$$Ax^{k+1} + Bx^k + Cx^{k-1} = g^{k+1}, \qquad k \geq 2 \tag{A.1}$$

has two (not necessarily distinct) nonzero real roots r_1 and r_2. Then, the generic solution to this equation has the form

$$x^\nu = c_1 r_1^\nu + c_2 r_2^\nu + \frac{1}{A} \sum_{l=2}^{\nu} r_1^{\nu-l} \sum_{s=2}^{l} r_2^{l-s} g^s, \quad c_1, c_2 \in \mathbb{R}.$$

Proof. It is sufficient to show that

$$\bar{x}^\nu = \frac{1}{A} \sum_{l=2}^{\nu} r_1^{\nu-l} \sum_{s=2}^{l} r_2^{l-s} g^s, \quad \nu \geq 2,$$

with $\bar{x}^1 = \bar{x}^0 = 0$ is a particular solution of (A.1).

Let $n \geq 1$. Multiply (A.1) by r_2^{2n-k-2} and add all the results for $k = 1, \ldots, n$. Setting $x^1 = x^0 = 0$, we obtain

$$Ar_2^{n-2}x^{n+1} + r_2^{n-2}(Ar_2 + B)x^n + \sum_{k=2}^{n-1} \left[(Ar_2^{2n-k-1} + Br_2^{2n-k-2} + Cr_2^{2n-k-3})x^k \right] = \sum_{s=2}^{n+1} r_2^{2n-s-1} g^s,$$

which, since r_2 is a root of the characteristic polynomial, implies

$$Ax^{n+1} + (Ar_2 + B)x^n = \sum_{s=2}^{n+1} r_2^{n+1-s} g^s, \quad n \geq 2. \tag{A.2}$$

Let $\nu \geq 1$. Multiply (A.2) by $r_1^{\nu-n}$ and add all the results for $n = 1, \ldots, \nu$. We obtain

$$Ax^{\nu+1} + \sum_{l=2}^{\nu} \left[r_1^{\nu-l} (A(r_1 + r_2) + B) x^l \right] = \sum_{l=2}^{\nu+1} r_1^{\nu+1-l} \sum_{s=2}^{l} r_2^{l-s} g^s, \quad \nu \geq 1.$$

Since r_1, r_2 are roots of the characteristic polynomial of the recursion equation, we have $B = -(r_1 + r_2)A$, which implies

$$x^{\nu+1} = \frac{1}{A} \sum_{l=2}^{\nu+1} r_1^{\nu+1-l} \sum_{s=2}^{l} r_2^{l-s} g^s, \quad \nu \geq 1.$$

Hence, $\bar{x}^\nu$ is a particular solution of (A.1). □

Proposition 12. *Assume that the coefficients of the three term recursion inequality*

$$Ay^{k+1} + By^k + Cy^{k-1} \leq g^{k+1}, \quad k \geq 1, \tag{A.3}$$

satisfy

$$A > 0, \quad C \geq 0, \quad A + B + C \leq 0.$$

Let $\{y^k\}_{k\geq 0}$ *be a solution to* (A.3) *with initial data* y^0 *and* y^1. *If* $\{x^k\}_{k\geq 0}$ *solves* (A.1) *with initial data* $x^0 = y^0$ *and* $x^1 = y^1$, *then the following estimate holds*

$$y^\nu \leq x^\nu, \quad \forall \nu \geq 0.$$

Proof. This is a comparison argument *à la* Grönwall. Let $\{z^k\}_{k\geq 0}$ be the sequence defined by $z^\nu = y^\nu - x^\nu$. Let us prove by induction that $z^k \leq z^{k-1}$, for all $k \geq 1$. The claim holds true for $k = 1$ since $0 = z^1 \leq z^0 = 0$. Assume now that $z^\nu \leq z^{\nu-1}$ for all

$1 \leq \nu \leq k$. The definition of $\{x^k\}_{k\geq 0}$ implies

$$Az^{k+1} + Bz^k + Cz^{k-1} \leq 0, \quad \forall k \geq 1.$$

Hence

$$Az^{k+1} \leq Az^k - (A + B + C)z^k + C(z^k - z^{k-1}) \leq Az^k,$$

which proves the claim. □

The following corollary is a specialization of the two previous results which will be needed in Section A of Chapter IV.

Corollary 6. *The three term recursion equation*

$$3x^{k+1} - 4x^k + x^{k-1} = g^{k+1}, \quad k \geq 1, \tag{A.4}$$

has the following general solution

$$x^\nu = c_1 + \frac{c_2}{3^\nu} + \sum_{l=2}^{\nu} \frac{1}{3^{\nu+1-l}} \sum_{s=2}^{l} g^s, \qquad c_1 \in \mathbb{R},\ c_2 \in \mathbb{R}.$$

Let $\{y^k\}_{k\geq 0}$ be the solution to the three term recursion inequality

$$3y^{k+1} - 4y^k + y^{k-1} \leq g^{k+1}, \quad k \geq 1,$$

with initial data y^0 and y^1. If $\{x^k\}_{k\geq 0}$ is the solution to (A.4) *with initial data $x^0 = y^0$ and $x^1 = y^1$, then the following estimate holds*

$$y^\nu \leq x^\nu, \quad \forall \nu \geq 0.$$

Proof. To obtain the generic solution, it is sufficient to notice that the roots of the characteristic polynomial of the equation are $r_2 = 1$ and $r_1 = 1/3$. To obtain the estimate it is sufficient to notice that $A = 3 > 0$, $C = 1 > 0$ and $A + B + C = 3 - 4 + 1 = 0 \leq 0$. □

VITA

Abner Jonatán Salgado González was born in Guatemala City, Guatemala. He obtained his B.S. and M.S. degree *with honors* in applied mathematics from Saint-Petersburg State Polytechnic University, Russia in June 2004 and May 2006, respectively. In August 2006 he began his Ph.D. studies in mathematics at Texas A&M University. During the summer of 2008 he worked as a summer intern at the *Laboratoire Jacques-Louis Lions* in Paris, France. He obtained his Ph.D. degree in August 2010.

Abner Salgado can be reached by writing to the Department of Mathematics, Texas A&M University, College Station, TX 77843-3368, or to the email address `abnersg@math.tamu.edu`.